Research Methods in Applied Linguistics

17

Research Synthesis Methodologies in Applied Linguistics

Edited by Sin Wang Chong

John Benjamins Publishing Company

Research Synthesis Methodologies in Applied Linguistics

Research Methods in Applied Linguistics (RMAL)

ISSN 2590-096X

The *Research Methods in Applied Linguistics* (RMAL) series publishes authoritative general guides and in-depth explorations of central research methodology concerns in the entire field of Applied Linguistics. The hallmark of the series is the contribution to stimulating and advancing professional methodological debates in the domain. Books published in the series (both authored and edited volumes) will be key resources for applied linguists (including established researchers and newcomers to the field) and an invaluable source for research methodology courses.

Main directions for the volumes in the series include (but are not limited to): Comprehensive introductions to research methods in Applied Linguistics (authoritative, introductions to domain-non specific methodologies); In-depth explorations of central methodological considerations and developments in specific areas of Applied Linguistics (authoritative treatments of domain-specific methodologies); Critical analyses that develop, expand, or challenge existing and/or novel methodological frameworks; In-depth reflections on central considerations in employing specific methodologies and/or addressing specific questions and problems in Applied Linguistics research; Authoritative accounts that foster improved understandings of the behind the scenes, inside story of the research process in Applied Linguistics.

For an overview of all books published in this series, please see *benjamins.com/catalog/rmal*

Editor

Rosa M. Manchón
University of Murcia

Editorial Board

David Britain
University of Bern

Gloria Corpas Pastor
University of Malaga

Marta González-Lloret
University of Hawai'i

Laura Gurzynski-Weiss
Indiana University Bloomington

Juan Manuel Hernández-Campoy
University of Murcia

Ute Knoch
University of Melbourne

Anthony J. Liddicoat
University of Warwick

Brian Paltridge
University of Sydney

Diane Pecorari
University of Leeds

Luke Plonsky
Northern Arizona University

Li Wei
University College London

Volume 17

Research Synthesis Methodologies in Applied Linguistics
Edited by Sin Wang Chong

Research Synthesis Methodologies in Applied Linguistics

Edited by

Sin Wang Chong
University of St Andrews

John Benjamins Publishing Company
Amsterdam / Philadelphia

The paper used in this publication meets the minimum requirements of the American National Standard for Information Sciences – Permanence of Paper for Printed Library Materials, ANSI Z39.48-1984.

DOI 10.1075/rmal.17

Cataloging-in-Publication Data available from Library of Congress:
LCCN 2026028389 (PRINT) / 2026028390 (E-BOOK)

ISBN 978 90 272 3769 9 (HB)
ISBN 978 90 272 3768 2 (PB)
ISBN 978 90 272 4353 9 (E-BOOK)

© 2026 – John Benjamins B.V.
This e-book is Open Access under a CC BY-NC-ND 4.0 license.
https://creativecommons.org/licenses/by-nc-nd/4.0

This license permits reuse, distribution and reproduction in any medium for non-commercial purposes, provided that the original author(s) and source are credited. Derivative works may not be distributed without prior permission.

This work may contain content reproduced under license from third parties. Permission to reproduce this third-party content must be obtained from these third parties directly.

Permission for any reuse beyond the scope of this license must be obtained from John Benjamins Publishing Company, rights@benjamins.nl

John Benjamins Publishing Company · https://benjamins.com

Table of contents

About the contributors ix

Preface xv
John M. Norris

SECTION 1. Introduction: The origin and evolution of research synthesis in Applied Linguistics

CHAPTER 1. The inconvenient truth about research syntheses in Applied Linguistics 2
Sin Wang Chong

SECTION 2. Research synthesis in Applied Linguistics journals

CHAPTER 2. Research syntheses in the international journal, *Language Testing*: Publishing trends and a drive to improve research reporting standards 22
Talia Isaacs, Ruslan Suvorov, Xun Yan & Sofya Styrina

CHAPTER 3. Research synthesis and meta-analysis in *SSLA*: Evidence, quality, and ways forward 41
Luke Plonsky, Yuhang Hu & Shujing Zhao

CHAPTER 4. Planning, doing, and disseminating research syntheses: An introduction to the new journal, *Research Synthesis in Applied Linguistics* 57
Sathena Chan & Sin Wang Chong

SECTION 3. Bibliometric/scientometric review

Section overview: What is a bibliometric/scientometric review? 68
Zhang Sai & Vahid Aryadoust

CHAPTER 5. Methodological considerations in conducting bibliometric/scientometric reviews 72
Peter Crosthwaite, Sulistya Ningrum & Jihua Dong

CHAPTER 6. Towards methodological rigor and analytical depth in conducting bibliometric reviews: A case study in Applied Linguistics 90
Zhang Sai & Vahid Aryadoust

SECTION 4. Meta-analysis
Section overview: What is a meta-analysis? **112**
Zhouhan Jin & Reza Norouzian

CHAPTER 7. Lessons from a modern meta-analysis: Methodological insights **115**
Zhouhan Jin & Reza Norouzian

CHAPTER 8. Learning through conducting meta-analyses: Supervisor and student reflections **134**
Stuart Webb & Su Kyung Kim

SECTION 5. Qualitative research synthesis and practice review
Section overview: What are qualitative research synthesis, qualitative synthesis of research, and practice review? **156**
Sho Kato

CHAPTER 9. Supervising master's students conducting qualitative research synthesis in a Japanese context **161**
Eoin Jordan, Jason Kushnir & Sho Kato

CHAPTER 10. Conducting a research synthesis on translanguaging practices in Anglophone classrooms **180**
Ming Sum Kong & Sarah Hopkyns

SECTION 6. Scoping review
Section overview: What is a scoping review? **202**
Mara Fuertes Gutiérrez & Inma Álvarez

CHAPTER 11. Reflecting on scoping review methodology with a language testing example **208**
Sha Liu

CHAPTER 12. Mapping research evidence with a scoping review: A critical appraisal **229**
Inma Álvarez & Mara Fuertes Gutiérrez

SECTION 7. Systematic review
Section overview: What is a systematic review? **248**
Hamish Chalmers

CHAPTER 13. Conducting systematic reviews in Applied Linguistics: A case study and critical reflection 253
Emily Oxley

CHAPTER 14. Challenges and practical solutions in Applied Linguistics systematic review: A reflective perspective 271
Takumi Aoyama

CHAPTER 15. Sing a song of synthesis: Harmony and discordances in preparing a systematic review on the effects of songs in second language education 292
Hamish Chalmers & Catherine Hamilton

SECTION 8. Emergent and interdisciplinary practices in research synthesis
CHAPTER 16. The opportunities and challenges of conducting a methodological synthesis on the applications of Q methodology in Applied Linguistics 311
Nicola Morea & Hessameddin Ghanbar

CHAPTER 17. Human reviewers in an AI era: A methodological reflection on research synthesis 330
Dongxia Nie, Qi Liu, Yixue Dominique Jiang, Rui Yan, Yi Zhang, Xin Fang, Jingying Hu, Xuechun Huang & Jing Yun

CHAPTER 18. Responding to the challenges of evidence synthesis: Some lessons from healthcare 348
Anastasios Bastouis, Burak Kundakci & Andrew Booth

SECTION 9. Research synthesis in applied linguistics: Looking ahead
CHAPTER 19. A principled approach to methodological agility in research synthesis: Putting ethics at the centre 368
Sin Wang Chong

Index 379

About the contributors

Inma Álvarez, Professor of Education at the School of Languages and Applied Linguistics at The Open University (UK). She was Director of Postgraduate Research Studies (2016–2025) for PhDs and Professional Doctorates in Health, Education and Applied Linguistics. She has published on the language learning experiences of doctoral students. https://orcid.org/0000-0002-0667-8264

Vahid Aryadoust is an Associate Professor of language assessment at the National Institute of Education of Nanyang Technological University, Singapore. His areas of interest include generative AI in language assessment, meta-analysis, and sensor technologies such as eye tracking, brain imaging, and GSR in language assessment.

Anastasios Bastounis is a Research Associate at the University of Sheffield (SCHARR). An epidemiologist and research methodologist, he specialises in evidence synthesis, big-data analysis, and meta-research, spanning behaviour change, mental health, infectious diseases, and applications of AI/LLMs in research. https://orcid.org/0000-0001-5861-9373

Andrew Booth is Professor in Evidence Synthesis at the Sheffield Centre for Health and Related Research (SCHARR) University of Sheffield. He is Co-Director of the Sheffield EnSygN Evidence Synthesis Group and a Co-Convenor of the Cochrane Qualitative and Implementation Methods Group. https://orcid.org/0000-0003-4808-3880

Hamish Chalmers is a Senior Lecturer in Applied Linguistics at the Department of Education, University of Oxford. His principal research interests are the education of multilingual learners in mainstream contexts, evidence synthesis, public engagement in science, and trials methodology. He is Editor of the International Database of Education Systematic Reviews (IDESR.org). https://orcid.org/0000-0003-2687-9722

Sathena Chan is Associate Professor of Language Assessment and Director of the Institute of Humanities, Education and Arts Research (IHEAR), University of Bedfordshire. She specialises in the design and validation of high-stakes English language assessment. She is President of UK Association for Language Testing and Assessment (UKALTA). She edits *Research Synthesis in Applied Linguistics*. https://orcid.org/0000-0002-7852-6737

Sin Wang Chong is Director of the Centre for International, Language and Teacher Education Research (CILTER), University of St Andrews. In a visiting capacity, he is Professor in Education at King's College London. He edits *Research Synthesis in Applied Linguistics* and *Review of Education*. https://orcid.org/0000-0002-4519-0544

Peter Crosthwaite is Associate Professor of Applied Linguistics at the School of Languages and Cultures, University of Queensland, Australia. He is a leading scholar on corpus linguistics, academic literacies, and second language writing, and is EIC of *Australian Review of Applied Linguistics*. https://orcid.org/0000-0002-1482-8381

Jihua Dong is Professor, Qilu Young Scholar, and Taishan Young Scholar at Shandong University, China. Her research interests include corpus linguistics, data-driven teaching, and academic writing. She has published in Applied Linguistics, English for Specific Purposes, Journal of English for Academic Purposes, ReCALL and System, among others. https://orcid.org/0000-0001-7864-2319

Xin Fang is currently a Research Fellow in the Department of Linguistics and Translation at City University of Hong Kong. She holds a PhD in Applied Linguistics from the University of Liverpool. Her research interests include second language acquisition (SLA), classroom interaction, and teacher talk. More recently, her work has extended to examining SLA within AI-mediated learning environments. In particular, she has incorporated artificial intelligence into project-based learning (PBL) in language education, and has designed and implemented AI literacy-oriented tasks to support learners' critical and pedagogically meaningful engagement with AI tools in PBL contexts. https://orcid.org/0000-0002-3046-3714

Mara Fuertes Gutiérrez, Senior Lecturer at The Open University (UK), specialises in Applied Linguistics, with a focus on Spanish Language Teaching, teacher education and inclusive languages curricula. Her recent works include *Technology-Mediated Language Teaching: From Social Justice to Artificial Intelligence* (Multilingual Matters, 2025). She is the Vice-President of ASELE. https://orcid.org/0000-0002-9890-5945

Catherine Hamilton is a researcher and lecturer in Quantitative Research Methods and EAL at the Department of Education, University of Oxford. Her research investigates the use of creative approaches for teaching foreign languages in early education, such as learning French through songs. https://orcid.org/0000-0001-7400-968X

Sarah Hopkyns is a lecturer at the University of St Andrews and UCL, UK. She has previously worked in the UAE, Canada, and Japan. Her research interests include language and identity, language policy, translingual practice, linguistic ethnography, linguistic landscapes, and English-medium instruction (EMI). She has published widely on these areas. https://orcid.org/0000-0003-0152-4108

Jingying Hu is a PhD candidate in Linguistics at Purdue University. Her research focuses on bilingual sentence processing, second language writing assessment, syntactic alternations, and information theory, using psycholinguistic experiments (e.g., self-paced reading, eye-tracking), corpus analysis, and computational modeling (natural language processing and machine learning). https://orcid.org/0009-0003-6582-1707

Yuhang Hu received her PhD in Applied Linguistics from Northern Arizona University. Her research interests include quantitative research methods, research synthesis, meta-analysis, and individual differences in second language acquisition.

Xuechun Huang is a PhD candidate at the School of Education, University of Leeds. Her research focuses on corpus linguistics, vocabulary teaching and learning, and translanguaging. She serves as the Events Team Lead for Open Applied Linguistics (AILA Research Network). Her work has been published in journals including *International Journal of TESOL Studies, Review of Education*, and *Languages*. https://orcid.org/0009-0004-8040-0836

Talia Isaacs is Professor of Applied Linguistics and TESOL at the UCL Institute of Education, University College London, where she currently serves as departmental Head of Research. She holds honorary status at Oxford University and has affiliation with Anaheim University. She is co-editor of the international journal, *Language Testing*. https://orcid.org/0000-0003-4302-3379

Dominique Yixue Jiang is a PhD student in Applied Linguistics at the UCL Institute of Education. Her research focuses on second language acquisition and assessment, with a particular interest in writing processes. She has contributed to research projects on neurocognitive processes underlying L2 writing and research synthesis across disciplines. https://orcid.org/0009-0009-0138-728X

Zhouhan Jin is an Assistant Professor at the University of Macau, China. Her research focuses on classroom-based language learning, particularly note-taking, vocabulary acquisition, and meta-analysis. Her work has been published in leading journals such as *Studies in Second Language Acquisition, Language Learning, The Modern Language Journal*, and *Applied Linguistics*. https://orcid.org/0000-0001-5285-9384

Eoin Jordan is a Professor within and Director of the International Education and Lifelong Learning Institute at University of St Andrews. His scholarly activities are focused on the enhancement of educational practice, often through the use of technology, but also based on an understanding of linguistic and cultural contexts. https://orcid.org/0000-0001-7382-0964

Sho Kato is a DProf TESOL student at the University of St Andrews and an English teacher at Kariyakita Senior High School in Aichi, Japan. His interests include vocabulary teaching and learning, self-regulated learning, language learner/teacher autonomy, technology, and research synthesis in applied linguistics. https://orcid.org/0009-0004-2021-4433

Ming Sum Kong is an English academic tutor at the University of East London. She is also a doctoral student in TESOL at the University of St Andrews. As a multilingual practitioner-researcher, her interests lie in language teaching and learning, translanguaging, multilingualism, and bridging the research-practice gap. https://orcid.org/0009-0001-8222-9022

Su Kyung Kim is a researcher at Korea University. Her research has focused on spacing effects in second language learning, individual difference factors for second language vocabulary, and meta-analytic approaches to second language research. Her broader research interests concern the processing, acquisition, and use of second language knowledge. https://orcid.org/0000-0001-6056-7760

Burak Kundakci is a Research Associate in Evidence Synthesis at the University of Sheffield and a Research Fellow at the University of Manchester. With a background in physiotherapy, his research focuses on evidence synthesis methodologies, including conventional and network meta-analysis, Delphi methodology, and outcome measure development in rheumatic/musculoskeletal diseases.

Qi Liu is a fully-funded PhD student at the University of Leicester. She is an Honorary Research Assistant at the University of St Andrews in Scotland. She is the Social Media Editor for the SSCI-indexed journal, Innovation in Language Learning and Teaching. https://orcid.org/0009-0005-7105-000X

Dongxia Nie is a PhD candidate at the Department of Education, University of Oxford and a researcher at the AI in Education at Oxford University (AIEOU) Hub. Her research focuses on digital language learning, second language (L2) reading for young learners, research synthesis in Applied Linguistics, and AI in education. https://orcid.org/0009-0006-4698-8407

Reza Norouzian is a Senior Researcher of Multilingualism at the Oregon Department of Education and serves on the Core Research Team. He sits on the editorial

boards of *Studies in Second Language Acquisition* and *Research Synthesis in Applied Linguistics*. His research focuses on instructed second language acquisition, meta-analysis, and research design, with publications in leading applied linguistics journals. https://orcid.org/0000-0002-0531-6886

Sulistya Ningrum is an English lecturer at Politeknik Negeri Sriwijaya, Palembang. Her research focuses on Translation, Corpus Linguistics, and Second Language Writing. She is also an award-winning fiction writer and currently develops creative writing projects and translation-based applications to support innovative and effective language learning. https://orcid.org/0000-0002-9938-9698

Nicola Morea is a Senior Impact Evaluator at Cambridge University Press & Assessment (Cambridge, United Kingdom). Nicola's research focuses on language education, teacher education, multilingualism, language testing, and research methods in applied linguistics and impact evaluation. https://orcid.org/0000-0003-0623-3078

Emily Oxley is a lecturer in Language and Literacy Education at the University of Glasgow. Her research focuses on children's language and literacy development and educational research methodologies including the Delphi consensus. She currently sits on the Editorial Board of *Research Synthesis in Applied Linguistics*. https://orcid.org/0000-0002-5736-9808

Luke Plonsky is Professor of Applied Linguistics at Northern Arizona University. His work, focusing primarily on SLA and research methodology, has appeared in over 150 articles, book chapters, and books. Luke serves as Editor of *SSLA*, Managing Editor of *Foreign Language Annals*, and Founding Editor of Applied Linguistics Press. https://orcid.org/0000-0002-5791-1839

Sofya Styrina is a PhD candidate in the Department of Linguistics at the University of Illinois Urbana-Champaign. Her research focuses on online discourse and the role of emotion in patient healthcare decisions. She currently serves as the editorial assistant for the journal *Language Testing*. https://orcid.org/0009-0001-0854-9853

Ruslan Suvorov is Associate Professor of Applied Linguistics at the University of Western Ontario, Canada. His research focuses on computer-assisted language learning and language assessment, with particular emphasis on second language listening and biometric methods. He currently serves as Associate Editor of the journal *Language Testing*. https://orcid.org/0000-0003-2736-5205

Stuart Webb is a Professor of Applied Linguistics at the University of Western Ontario. Before teaching applied linguistics, he taught English as a foreign lan-

guage in Japan and China for many years. His research interests include vocabulary studies, second language acquisition, and extensive reading, listening, and viewing. His latest books are How Vocabulary is Learned (with Paul Nation) and The Routledge Handbook of Vocabulary Studies. https://orcid.org/0000-0002-8297-4997

Sai Zhang is a lecturer at the School of Foreign Studies, Hebei Normal University, China. Her research interests include research synthesis, research evaluation, language assessment, and translation. Her work has been published in leading journals, including *Studies in Second Language Acquisition* and *Language Testing*. https://orcid.org/0009-0001-7732-5618

Yi Zhang received her PhD in Applied Linguistics from the University of Warwick. Her research interests include classroom interaction, discourse analysis, language teacher education, and ELT textbook development and use.

Shujing Zhao is a Ph.D. student in Applied Linguistics at Northern Arizona University. She currently serves as an editorial assistant for *Studies in Second Language Acquisition*. Her research focuses on individual differences, Task-Based Language Learning (TBLT), and research methods. https://orcid.org/0009-0006-6726-6980

Xun Yan is Professor of Linguistics, Second Language Acquisition and Teacher Education, and Educational Psychology at the University of Illinois Urbana-Champaign. His research focuses on speaking and writing assessment, psycholinguistic and computational approaches to language testing, and language assessment literacy. He is co-editor for the journal *Language Testing*. https://orcid.org/0000-0002-4544-4938

Jing Yun, a DProf TESOL student at St Andrews University. She obtained an MA in Applied Linguistics with TESOL at Sheffield University, an MA in Cultural Studies at KU Leuven. Jing serves as Research Editor of *St Andrews Journal of International and Language Education*. Her research focuses on language advising and research synthesis. https://orcid.org/0000-0002-4573-8522

Rui Yan holds a PhD in Education at the University of Bath, UK. Her research interests include educational technology, computer-assisted language learning, and gamification. Her work has been presented at international conferences including AAAL, EuroSLA, EuroCALL, and BAAL. She has contributed to publications such as *The Palgrave Encyclopedia of Computer-Assisted Language Learning* and the *Routledge Handbook of Technological Advances in Researching Language Learning*. https://orcid.org/0009-0005-7278-7396

Preface

John M. Norris
Educational Testing Service Japan

Several decades have passed since Lourdes Ortega and I challenged the field(s) of applied linguistics to adopt a synthetic ethic for guiding and improving both primary and secondary research practices. Our goals at the time were straightforward: (a) to encourage the adoption by primary researchers of rigorous practices for situating and reporting their work in relation to existing research, and (b) to expand the application of rigorous secondary reviews for systematically and comprehensively assessing the state of accumulated knowledge within specific research domains. It is particularly satisfying to see this new volume, *Research synthesis methodologies in applied linguistics*, appear now, twenty years after the publication of *Synthesizing research on language learning and teaching*, with both volumes published under the John Benjamins label. The subject matter and contributions in this new volume, masterfully compiled by editor Sin Wang Chong, demonstrate resoundingly the extent to which research synthesis has become a critically important practice across a variety of applied linguistics domains in the interim. Here, by way of situating the current volume in relation to some of that trajectory, I highlight a few observations on the impact of a synthetic ethic in applied linguistics research, the ways in which the chapters in this book exemplify that ethic, and a handful of current and future challenges.

A clear contribution of research synthesis has been the improvement in primary study research methodologies and reporting practices across multiple applied linguistics domains. Most indicatively, many scholarly journals have adopted robust standards for the reporting of primary research, with a particular emphasis on the essential elements of quantitative studies, from the treatment of descriptive and inferential statistics to the inclusion and interpretation of effect sizes and confidence intervals. Considerable attention has also been paid to statistical reform and the nuanced use of null hypothesis statistical testing, resulting from gaps in interpretation that were first laid bare in early applied linguistics meta-analyses. Along the way, critical methodological reviews have focused primary researchers on challenges in study design and analysis, and validity generalization studies have inspired the development of new and better-fitting measures of key dependent variables. Likewise, the inherent value of replication — of vari-

ables, designs, measures — has received considerable attention, leading to a meaningful focus on repeated investigation of the factors that really matter in various research domains. In short, primary research has evolved in notably positive ways in response to findings and challenges provided by applied linguistics research syntheses, and chapters in the first two sections of this book do a good job of highlighting examples of this impact in specific fields of endeavor.

In a similar vein, it is particularly encouraging that quite a few applied linguistics scholarly journals have adopted dedicated sections and/or regular special issues devoted to reporting research syntheses of various kinds. Not only does this practice underscore the critically important role played by synthesis in advancing scientific knowledge accumulation and research innovation, but it also helps establish domain-specific synthetic methodologies, criteria, and frames of reference, thereby tailoring and enhancing its value. Section 2 of the book offers insightful reflections on these developments from the perspective of several such journals. Of note, there is also a chapter introducing a new journal devoted entirely to, per the title, *Research Synthesis in Applied Linguistics*. The advent of a new journal is testament to the increasing and widespread uptake of synthetic methods for addressing core problems across this diverse field of endeavors; indeed, a journal devoted to a research methodology is a rarity in applied linguistics.

Another bright spot in the on-going development of research synthesis in applied linguistics over the past several decades has been the exploration, application, and innovation of new and diverse synthetic approaches and methods. While compiling the contributions for *Synthesizing research on language learning and teaching* in the early 2000s, we were hard pressed to find examples (one each) of historiographic and qualitative syntheses in addition to a handful of meta-analyses. Much has happened in the interim by way of increased sophistication and diversification of synthetic approaches and methods. Methods for meta-analysis have been fine-tuned to the particular challenges of quantitative applied linguistics research, including for example the development of field-specific benchmarks for interpreting effect sizes and the implementation of hierarchical and multilevel models for analyzing non-independent data (a major challenge for quasi-experimental research traditions). Beyond meta-analysis, diverse approaches have been adopted to better address a variety of specific questions of interest from a synthetic perspective. These include, for example, methodological reviews that focus on patterns and problems in research methods for specific domains, bibliometric reviews of publishing practices and related phenomena, qualitative syntheses for ethnographic and related research, and systematic, narrative, and scoping reviews, among others. Sections 3 through 8 of this volume provide recent examples and insightful discussion of these approaches to broadening the potential for synthetic perspectives on applied linguistics research. Of

particular interest to me is the attention paid to scoping reviews, an approach that may be especially well suited to domains of research that are complex, contentious, and multimethodological (i.e., much of the research in applied linguistics), and one that my co-authors (Shoko Sasayama, Aleksandra Malicka) and I recently adopted in *Second language cognitive task complexity: A research synthesis*, also published by John Benjamins.

Of course, research syntheses of various kinds have also led to important, often major, iterations in substantive findings and the state of knowledge in quite a few distinct domains of applied linguistics research, as reflected in several chapters in this book. In particular, intervention-focused research (e.g., on instructional effectiveness, learner motivation, technology-mediated teaching and learning) has benefited from synthetic findings. At the same time, synthetic methodologies continue evolving to accommodate distinct types of primary research, including for example meta-evaluation of program-level evaluation studies and practice-oriented reviews with a specific orientation to practitioner concerns. These and other innovations, including the adoption of explicitly interdisciplinary approaches to synthesis, are reflected in the final sections of the book.

Though much has been accomplished in pursuit of the synthetic ethic in applied linguistics, challenges remain, as addressed in the first and other chapters throughout the book. Without a doubt, there has been a concerning tendency to over-apply meta-analytic techniques to immature domains, coupled with the over-interpretation of findings as presenting conclusive evidence without attending to a variety of moderating factors, something we warned against already in the early 2000s. The increased availability of automated approaches to calculating and combining effect sizes, while making meta-analysis more accessible to broader audiences of researchers, may also have contributed to the dramatic increase in premature or relatively uninformed meta-analysis. The high frequency of meta-analyses being reported, relative to other synthetic options, may also reflect a lack of methodological sophistication within applied linguistics regarding the most fitting synthetic approach for addressing a given problem. By presenting a host of diverse methodologies for research synthesis, this book goes a long way towards addressing some of these issues. Additionally, it is worth emphasizing the important contributions to be made by involving substantive domain experts in the synthetic enterprise as well as increasing collaborative efforts. While the individual expert narrative review may lean towards the selective or idiosyncratic representation of a domain, experts, ideally representing multiple salient perspectives, should play a vital role in guiding the focus and conduct of research synthesis as well as the interpretation of findings. International collaborative research networks may offer an important way forward in this regard, providing a kind of scholarly sounding board to facilitate meaningful syntheses.

Another more recent challenge for the contribution of research synthesis to applied linguistics bears careful attention as well. In an era where AI seems capable of doing just about anything, a note of caution is in order. Of course, AI systems have come to play a default role as 'synthesis' machines, automatically (and extremely quickly) scrubbing the internet for any ideas to inform an apparently authoritative response to whatever query a human might come up with. On the one hand, this development has placed a type of synthesis in the hands of the common user of the internet, without doubt a useful and hugely influential extension of the methodology. On the other hand, the AI approach to synthesis is somewhat concerning on several fronts. For one, the tendency of AI systems to hallucinate (e.g., about research publications and citations and authorship, case law, scientific findings and breakthroughs) runs counter to the core synthetic ethic of reporting what the actual accumulated evidence has to say — synthesis is fundamentally a factual, comprehensive, truth-seeking endeavor. Another concern is the drive (almost an overwhelming desire) of AI to produce an answer, whether or not there is an evidence-based answer to be found in the existing research. Yet, increasingly, humans enthusiastically assume and trust that AI will tell them what they want or need to know. It is even deceptively easy to envision the (simplistic) pursuit of (faux) scholarly research syntheses on the basis of AI systems, ceding control over search and retrieval, study coding, effect size calculation and comparison, and reporting of findings to the automated machines.

There is little doubt that AI is playing and will continue to play a role in the conduct of scholarly research synthesis in applied linguistics and other fields. It is therefore of even greater importance that human experts oversee all aspects of the synthetic enterprise. Given the vagaries of AI systems, human experts must remain firmly in the driver's seat, rendering decisions regarding what topics to synthesize in what ways, cross-checking facts and findings, and making interpretations about the accumulated evidence. To cede control of the enterprise to AI runs counter to the synthetic ethic.

The past twenty-plus years of research synthesis in applied linguistics has witnessed the rapid rise and diversification of possible approaches to and methods for making sense of accumulated research. It has also raised a handful of new challenges, where the goal of the synthetic ethic is the generation of trustworthy interpretations about the state of knowledge and how it might best be advanced. While synthesis has been embraced enthusiastically and applied vigorously in some quarters of applied linguistics, its development has not followed a particularly systematic trajectory, as signaled by some of the imbalances in reporting of syntheses raised prior. This first era of research synthesis might even be characterized as primarily exploratory in nature. In this regard, the final chapter of this book introduces a potentially quite useful way forward under the guidance

of a 'theory-of-change' approach. By overlaying a system for identifying the kinds of questions that merit synthetic attention and then reverse-engineering the fitting approach and methodology, this proposal may serve to propel the synthetic ethic into a new era. Where the outcome is a more nuanced, accurate, and helpful understanding of the state of knowledge in applied linguistics, then the on-going enterprise and related innovations will have been worth the effort.

SECTION 1

Introduction

The origin and evolution of research synthesis in Applied Linguistics

CHAPTER 1

The inconvenient truth about research syntheses in Applied Linguistics

Sin Wang Chong
University of St Andrews

In this chapter, I explain why this edited book on research syntheses in Applied Linguistics is needed and how this book contributes to the field by offering a "behind-the-scenes" perspective towards the ethical and methodological considerations before and while doing a research synthesis. Unconventionally, I begin this book, which promotes research syntheses, with a cautious tone and strong caveats about what research syntheses are and can do. I first present misconceptions about research syntheses and highlight the usefulness and limitations of synthetic research, focusing more on the latter. Then, I reflect on the impetus for putting together this book in view of the latest developments of research syntheses in Applied Linguistics. This is followed by an outline of book sections and chapters.

Keywords: research synthesis, synthetic research, applied linguistics, research ethics, methodological rigour, study quality, meta-research

The "turns" in Applied Linguistics

As a relatively young field of inquiry, Applied Linguistics has developed rapidly, shedding important and timely light on the multifaceted roles of how languages are learned and used in societies, communities, and professions. Responding to the evolving nature of languages and how meaning is constructed in an increasingly connected world, applied linguists have investigated a wide range of language-related topics, drawing liberally from a plethora of philosophical, epistemological, theoretical, and methodological traditions straddling disciplinary boundaries. To date, there have been a number of watershed moments in Applied Linguistics, also known as "turns". These "turns" signify an emergent or renewed focus on a topic, either substantive or methodological, which has resulted in a surge of research endeavours that propel the field toward a particular agenda. While there is no consensus on what these "turns" are, they can be

https://doi.org/10.1075/rmal.17.01cho
Available under the CC BY-NC-ND 4.0 license.© 2026 John Benjamins Publishing Company

referred broadly as the cognitive turn, the social turn, the multilingual turn, the psychological turn, and the methodological turn.

The *cognitive turn* focuses on an individual's cognitive capacity and processes that influence how languages are acquired, learned, and used (e.g., Krashen's Monitor Hypothesis). Hypotheses related to input, interaction, and output, and other cognitive science theories, have become foundational tenets in Applied Linguistics that are widely discussed, alongside their limitations. Linguistic accuracy, internal mental processes, representations, and mechanisms are the predominant focus of investigations in the cognitive turn. The *social turn* (e.g., Block, 2003) recognises language use does not exist within an individual or in a vacuum; applied linguists, influenced by Vygotskyan perspectives and others, postulate that language use is socially constructed (Lantolf & Pavlenko, 1995). Drawing on theoretical frameworks such as sociocultural theory, activity theory, ecological systems theory, and complex dynamic systems theory, the use of languages is examined in light of a myriad of contextual and sociocultural influences; other topics such as language socialisation and linguistic ethnography are explored.

The multilingual turn is one of the more recent waves of scholarly enthusiasm that has gained a lot of traction. Building on the social turn that recognises language as a social construct, some applied linguists advocate for the plurality of language, emphasising more localised use of languages, or varieties of a language. Also associated with the multilingual turn is an agenda of social justice (e.g., Cushing, 2024), challenging established norms such as native-speakerism and pushing for decolonisation of languages and how they are taught. The multilingual turn exerts influences far beyond academia, reaching professions, mostly in education, regarding how languages are learned and assessed. Translingual practices, including translanguaging, which encourages the fluid use of target and first languages, have the attention of many teaching professionals worldwide (Canagarajah, 2013; García & Li, 2014). The multilingual turn is also reflected in some academic journals in Applied Linguistics, which accept the use of authors' first language in the titles and abstracts of their publications.

The psychological turn is most widely known for its (renewed) emphasis on some individual learner differences, especially those pertaining to emotions (e.g., Dörnyei & Ryan, 2015; MacIntyre et al., 2016). Representative research of the psychological turn draws on positive psychology and educational psychology to unravel how the psycho-affective filter influences how languages are learned and used. In particular, a wide range of emotions has been studied including grit, pride, boredom, and enjoyment.

While the above agendas are substantive in focus, *the methodological turn* in Applied Linguistics (e.g., Byrnes, 2013) asserts the need to focus on meta-science, which refers to the science of science, or to put simply, the study of how

research in Applied Linguistics is conducted and how study quality is understood, including methodological and ethical considerations. These are exemplified by book series such as Research Methods in Applied Linguistics published by John Benjamins, edited books (e.g., De Costa, 2015; De Costa et al., 2024), special issues published in international journals (e.g., De Costa & Isbell, 2024), and the founding of the first methodological journal in our field, *Research Methods in Applied Linguistics*. There is also the open science movement that promotes transparent and reproducible research practices, as well as facilitates the uptake of research through effective science communication strategies targeting a non-academic audience (Liu et al., 2023; Plonsky, 2024a).

These "turns" witness the development of Applied Linguistics — as "archipelagos" of distinct research interests and areas. While vibrant in their own right, these "turns" often operate in silos, forming their own closed-group communities. If Applied Linguistics were to truly progress and advance as an academic field, it is of paramount these "research islands" need to be in more active dialogue. Applied linguists need to take stock of key lessons learned from each sub-field and Applied Linguistics as a whole, avoid research waste (i.e., duplicated and ill-conceived research attempts) (Isaacs & Chalmers, 2025), and identify future research directions that bring value to society (Plonsky, 2024b). This is where research syntheses come in.

Research synthesis and research synthesist

Research synthesis refers to a survey of research evidence and research-related information that is systematic, transparent, and often pre-registered. Research synthesis addresses a broad range of topics, both substantive and methodological, using a range of methodological designs. As such, research synthesis is used in this book as an umbrella term to refer to these designs, including scoping review, systematic literature review, methodological synthesis, bibliometric review, meta-analysis, qualitative research synthesis, meta-review, meta-ethnography, practice review, and many more. In disciplines such as Medicine and Healthcare, the term "evidence synthesis" is usually used in lieu of "research synthesis". For this edited book, *research synthesis* is the preferred term. This may seem like a misguided decision given how well-established the term "evidence synthesis" is in other disciplines; in addition to reasons I outlined elsewhere (Chong, Bond, & Chalmers, 2024) (e.g., how the term was first used in Norris and Ortega, 2006, the first book-form publication on synthetic research in Applied Linguistics), I find the term "research synthesis" more encompassing than "evidence synthesis" because the former denotes the synthesis of not only research evidence or findings but also

other types of information commonly included in academic publications such as research methods, interventions, and citations. Such information is not "evidence" per se but useful insights that are analysed in some review types such as methodological syntheses, practice reviews, and bibliometric reviews.

Applied Linguistics has no shortage of advocates or proponents, those who preach innovative and transformational ideas, encouraging us to think ahead of our times; however, synthesists are much needed in Applied Linguistics. Synthesists act as observers or curators, signposting others to collective insights from a body of knowledge. Synthesists are also gatekeepers; they appraise methodological rigour of published studies and help others understand the strength of evidence in a given area. Synthesising research is not just a matter of looking back and taking stock; it is about paving the way for the future of Applied Linguistics. Synthesists are a rare breed in Applied Linguistics because they need specialised methodological training and are well-versed in a wide range of research methodologies. They are also adept at working in diverse teams, in close collaboration with researchers specialising in different sub-domains of Applied Linguistics. Due to the rarity, the works of a synthesist are not widely understood and sometimes under-appreciated.

On the other end of the spectrum, there are unsubstantiated claims boasting the values of research synthesis, especially in light of its elevated status in other disciplines. In disciplines such as Medicine, research syntheses are viewed as the crème de la crème of research due to their rigour and scope, with some even placing it above randomised controlled trials. In Applied Linguistics, there is a general consensus that research syntheses are substantial and influential pieces that set the tone of a particular area of research; they are also usually highly cited such as those published in *Language Teaching*, a leading journal in our field that has a long history of publishing high-quality research syntheses.

To fully appreciate the values of research synthesis, I find it important to debunk some of the misconceptions, which I will call "myths", about its design, implementation, and use. So, despite being slightly unorthodox, I choose to begin a book about research synthesis by unpacking its limitations and challenges. I will then share how the idea of a book on research synthesis came to fruition and how I envision it to contribute to the field of Applied Linguistics.

Myths about research syntheses

Myth #1: Research syntheses are "low-effort" research

I have come across this comment many times: "It is easy to do a research synthesis; it is like doing an annotated bibliography". With the rise of generative AI, a similar remark is made, questioning the need for researchers to engage in a synthesis as there are AI tools that can produce a report of key research insights on a topic in minutes. From my experience supervising master's students, some see doing a research synthesis as a shortcut to completing their dissertations because there is no need for ethical approval, and only basic research skills such as searching for and summarising articles are required.

None of this is true. Research synthesis is a summary of research insights. It is similar to a literature review in many ways but it is more than a succinct description of publications. If we examine the origin of the word "synthesis", "to synthesise" refers to putting parts together to form a coherent whole (Etymonline, 2025a). This is a much higher-order thinking process than summarising, which means to provide headlines of insights without the understanding of how these major discoveries connect, and how they help understand the area of inquiry. "To synthesise" also contrasts with the verb, "to analyse", which refers to the dissecting of a topic into discrete elements, examining one element at a time in-depth (Etymonline, 2025b) — this is what primary research is about. A high-quality research synthesis does more than just summarising findings; it is thorough, in-depth, and offers original knowledge and new understanding of a topic. This often involves sophisticated analysis of a large amount of information, deep insider's knowledge about an area, and effective communication skills that turn an ambitious scope of work into an accessible report.

Research syntheses are complex. To conduct a comprehensive search on a topic, researchers need access to a wide range of academic databases and research repositories. A lot of time needs to be spent on developing and refining the search strategy, screening records, extracting and synthesising insights. Research syntheses are also expensive to conduct because specialised software may be needed to conduct advanced analyses and facilitate large-group collaboration.

Myth #2: There is always a need for new research syntheses

Research syntheses are resource-intensive. It requires a dedicated team and a lot of their time. Therefore, it is essential to contemplate whether there is a real need for a new research synthesis and how it contributes to both research and practice. Based on an ongoing meta-review on Applied Linguistics that I am lead-

ing, the largest research synthesis to date in the field, we identified 1,736 research syntheses, covering virtually all areas of Applied Linguistics including language and technology, language pedagogy, language acquisition, language assessment, instrument use of language, interactional use of language, representational use of language, language-related construct, and meta-research (research about research) (Chong et al., 2025). It has become more difficult to identify areas for new research syntheses.

There are a few ways to help researchers identify new contributions of a research synthesis idea. First, researchers can apply a "small lens" approach to conducting research synthesis. A "small lens" approach to research, originally proposed by Ushioda (2016), focuses on the values of *in situ* research on language learning motivation, arguing that the investigation into contextualised knowledge can shed important light on influencing language teaching, learning, and policy. The same is true for research synthesis. While it is no exaggeration to say that there is at least one research synthesis on any given topic in Applied Linguistics, new research syntheses can adopt a "small lens" approach such as focusing on the same subject but on different educational contexts and levels, synthesising evidence published in different languages especially around regional topics, and presenting an overview that synthesises information from academic and non-academic sources (e.g., policy documents).

A second strategy to ensure originality of a research synthesis is to focus on evidence not already synthesised such as conducting a qualitative research synthesis to complement an existing meta-analysis that focuses on quantitative evidence. Third, consider updating a published review by replicating the search strategy and other methodological information. This is particularly useful for topics that are emergent and vibrant, such as those related to educational technology.

Regardless of how you intend to frame your research synthesis, it is crucial that researchers conduct preliminary searches to identify if there are similar reviews on the topic of interest or a cognate topic to look for an area that warrants further investigation. Platforms that host a plethora of published research syntheses or protocols of research syntheses include the international journal, *Research Synthesis in Applied Linguistics*, the *Research Synthesis in Applied Linguistics Special Interest Group* of the British Association for Applied Linguistics, and the *International Database of Education Systematic Reviews* (IDESR) managed by the University of Oxford. Additionally, researchers need to be aware that other reasons may make it unsuitable to progress a research synthesis idea such as the lack of primary studies on a topic.

Myth #3: Research syntheses are reliable

Research syntheses are highly influential publications. They are frequently cited because of their breadth, depth, and agenda-setting nature; they are also touted for being objective, which enhances the reliability and trustworthiness of such research. Unlike traditional literature reviews such as a narrative review, that rely on the authors' knowledge, background, and experience in selecting and discussing a focused sample of publications that are regarded as representative to the area of interest, research syntheses follow a series of systematic, predetermined, and transparent methodological steps that allows researchers to identify not only a representative sample but virtually all relevant publications. These steps usually include setting research questions, determining the search strategy, screening and deduplicating records, extracting data and information, appraising study quality, synthesising insights, and drawing implications for future research agendas in the domain.

The objectivity of a research synthesis rests on the following mechanisms. First, the transparency in the search strategy. This includes a thorough documentation of all the databases where the search is conducted and the complete search string(s) used for each search. Second, the inclusion and exclusion criteria contribute to the objectivity of a research synthesis. Instead of relying on an individual's judgment that deems one publication relevant to a review and another not, researchers doing a research synthesis are required to develop criteria that set the bar for including and excluding publications. These criteria usually cover areas such as timeframe, language of publication, type of publication, and other topic-specific considerations. However, it must be acknowledged that even with these criteria, there is a possibility that biases are introduced such as publication bias and selection bias (see myth #4). Third, the involvement of a second reviewer is required in different methodological stages including screening, data extraction, study appraisal, and data synthesis. A second reviewer is needed when screening is conducted, including title and abstract screening, and full-text screening but particularly important in the latter. A second reviewer is involved in piloting the data extraction form and comparing extracted information. The additional reviewer plays a similar role in verifying and confirming the accuracy when study quality is evaluated and when extracted information is synthesised. In the screening, data extraction, study appraisal, and data synthesis stages, inter-coder reliability is expected to be calculated and reported. There are different ways to calculate inter-coder reliability, but it is most commonly done via percentage or kappa, although there are pros and cons of these approaches.

There are publications out there that are called "research synthesis" without incorporating these objectivity measures. This "jingle fallacy" refers to "the erro-

neous belief that two different things are the same because they are given the same name" (Littell, 2018, p. 5). Researchers must be literate enough to identify signs that lead them to question the reliability and objectivity of a research synthesis, which include partial reporting of the search strategy, the lack of a second reviewer, and, in some cases, the omission of some of the fundamental methodological stages in the review. For instance, in a methodological synthesis focusing on qualitative research syntheses on computer-assisted language learning (Chong & Reinders, 2021), it was found that most of the included reviews omitted the data synthesis stage or mistakenly referred to it as synonymous with data extraction.

While objectivity and reliability are valuable characteristics of research syntheses that contribute to their credibility, they should be considered alongside reflexivity. Reflexivity in research refers to researchers' awareness of how their methodological decisions are shaped by various influences. Such hidden influences include personal experience, research preferences, fatigue, power dynamics, and timeline pressure. The goal of incorporating a reflexivity dimension in methodological reporting is to ensure these influences are visible and discussable, and to emphasise the "humanity" of researchers, celebrating their expertise and challenges (Consoli & Ganassin, 2025, p. 1). Reflexivity can be documented through the use of a researcher's logbook that presents an audit trail of what decisions are made, how they are made, and the process of reaching those decisions (Liu & Chong, 2024).

Myth #4: Research ethics does not apply to research syntheses

It is a common misconception that research ethics does not apply to research synthesis. This is true if you limit the idea of research ethics to applying for ethics approval. Yes, conducting a research synthesis does not normally require formal ethics approval from your institution (although some universities still require an expedited ethics application or a notification); however, research ethics means much more than completing an ethics application form and asking for informed consent from participants. Research ethics is about respect, care, and trust for "people; knowledge; the quality of… research; the environment; and academic freedom" (British Educational Research Association (BERA), 2024, p. 9). Since research synthesis is a form of research, the same ethical principles apply.

But what specifically is at stake if research ethics is not considered in a research synthesis? Like all kinds of research, different types of biases can be introduced in the process of conducting a research synthesis. Being aware of and guarding against these biases is especially important to those conducting a research synthesis because of the nature and influence of this type of research. Types of biases that are commonly found in research syntheses include: publi-

cation bias, selection bias, language bias, citation bias, and data extraction bias. *Publication bias* is the tendency for studies with positive or significant results to be published more often than those with null or negative results. *Selection bias* occurs when included studies in a research synthesis are not representative due to non-random selection; this can happen when researchers choose to focus on a specific publication type such as journal articles. *Language bias*, one that is particularly relevant to Applied Linguistics, refers to limiting inclusion to studies published in certain languages, typically English, which can exclude relevant studies. *Citation bias* is similar to publication bias because of the increased likelihood for research syntheses to include more highly cited papers, especially in the present time when AI-powered search engines like Litmaps and Semantic Scholar prioritise showing publications based on the number of citations they have received. While the above types of biases happen during the searching and screening stage, *data extraction bias* happens when there is inaccurate or inconsistent charting of evidence or information from included studies.

The dismissal of ethical considerations in research synthesis becomes obsolete given the increasing use of AI as a research tool. AI-supported tools and platforms for conducting research syntheses skyrocketed in recent years; some influential examples include Elicit, which automates the extraction of key information and data from publications, and Rayyan, which introduces new deduplication and screening features using AI functions. Even specialised software for conducting research syntheses like EPPI-Reviewer has incorporated large language models in its recent upgrades. While the use of AI in research syntheses is not the focus of this chapter, how AI is used to (semi-)automate the synthesis process can be broadly categorised into three types: exploration, expedition, and endorsement. *Exploration* refers to using AI to brainstorm or refine ideas; *expedition* means using AI to accelerate or support human-led processes such as using AI to help speed up deduplication; *endorsement* describes scenarios where AI helps generate conclusions, usually without adequate human oversight. Based on research evidence that validates AI use in research syntheses in other disciplines, the consensus is that AI performs better and is more useful in the earlier methodological stages of a research synthesis such as screening (Ruan et al., 2025), while its reliability is still questionable for later and more complex stages like study appraisal (Landerholm, 2025).

Returning to BERA's definition of research ethics (BERA, 2024), there is also a social justice angle to ethical research practices that are relevant to research syntheses. With the given research landscape in Applied Linguistics, most studies are still published primarily in the English language and are about the Global North, particularly the Anglosphere. While some of these are gradually shifting, researchers need to be aware of these systemic biases that give power and voice to

a selected and privileged group. Without careful consideration, we risk perpetuating such partiality. For instance, when researchers justify their inclusion of only journal articles published in the Social Sciences Citation (SSCI) Index by equating publications in SSCI-indexed journals as quality-assured outputs. While in most cases this is true, SSCI-indexed journals usually target an international audience, favouring topics that are Anglocentric. There are regional journals that publish equally rigorous and important works that are not included in SSCI simply because the journals focus on a local readership. Even quality indicators of research syntheses that we take for granted such as the inclusion of multiple research databases deprive researchers in low-resource settings of publishing research syntheses. At the time of writing this chapter, I had the opportunity to have a conversation with a researcher based in Vietnam, who is doing her PhD in Hong Kong, she expressed her worries that she will no longer have access to research databases that require an expensive subscription such as Web of Science after she graduates. Researchers must constantly challenge their methodological decisions from a social justice perspective; this is equally true, if not more so, for synthesists.

Myth #5: Research syntheses are useful for researchers only

Research syntheses give researchers a systematic view of where a field currently stands and where it is heading; at the same time, they point to future research directions. Researchers often refer to research syntheses to inform their research programmes and designs because synthetic research provides insights into areas that are under-researched through a systematic and methodologically rigorous survey of the literature. There are several ways that research syntheses are useful for researchers. Research syntheses, which summarise a large number of research studies on a topic, are used by researchers to develop conceptual frameworks for their studies. This is especially useful when researchers are exploring new concepts or notions, or those that are contested and rapidly developing. By using findings from research syntheses, researchers can develop more holistic, nuanced, and literature-informed frameworks for analysing their data. Another way that research syntheses are used by researchers is through reviewing existing research methods and instruments used to investigate a topic. This enables researchers to justify their choice of methods and instruments as well as adapt existing instruments for their research. Additionally, researchers conduct "living" research syntheses, syntheses that are regularly updated, to include the latest research. These "living" research syntheses function as research libraries that curate and categorise publications on a topic of interest. These databases are usually open to other researchers to browse and contribute to; researchers can also conduct further

analysis on published data. A living meta-review on Applied Linguistics research syntheses is one such example (Chong et al., 2025). This project is built on a meta-review, or a review of reviews, commissioned by the British Association for Applied Linguistics that surveys and analyses topics and methodological rigour of published research syntheses in our field. Using EPPI Reviewer and EPPI Visualiser, a living research synthesis database is being built that will enable applied linguists to examine research syntheses by topics, review types, and quality. The included research syntheses can be used by other researchers to conduct further analysis.

Beyond their use within research communities, research syntheses play a crucial role in bridging the gap between research and practice. Drawing on the concept of evidence ecosystem by Shepherd (2014), research evidence, especially that in applied fields of study such as Applied Linguistics, can only make a real impact on society if it is used by practitioners, policymakers, and those beyond academia. The concept of evidence ecosystem points to the need to consider not only evidence generation but also evidence translation that leads to evidence use. For most researchers, their attention focuses on generating new insights, sometimes at the expense of devising effective strategies to communicate their research to a broader audience. To make research evidence "land", evidence translation is a key step.

Evidence translation refers to the use of strategies, platforms, or mechanisms to communicate research findings in accessible ways. There are numerous successful examples in Applied Linguistics such as OASIS, which summarises research in plain language (Alferink & Marsden, 2023); TESOLgraphics, a platform that produces one-page infographics on published research syntheses and evidence-informed teaching materials (Sato et al., 2024); and there is the TBLT Language Learning Task Bank (Gurzynski-Weiss & IATBLT, n.d.). Despite their success, these initiatives are researcher-led, and encouraging practitioners and policymakers to engage with research publications remains challenging. One major reason for such a lack of research engagement is time (Chong, 2020). Reading a journal publication takes time and keeping abreast of the latest trends and findings on a topic takes not only dedication but access and resources that people outside of academia rarely have. Research syntheses can be the solution. Instead of asking practitioners and policymakers to read multiple publications on a topic, a research synthesis provides an overview of the latest evidence and description of practice on a topic of interest. Although it is usually longer than a primary study, a research synthesis is easier to read because it contains visual aids that summarise large amount of information such as tables, diagrams, and figures; there are also appendices with additional information. Because research syntheses are

frequently commissioned or funded by research councils or government bodies, they are more often freely available than primary studies.

One type of research syntheses that is relevant to practitioners and policymakers is practice reviews. An emergent form of research synthesis in Applied Linguistics, a practice review is widely used in educational research in certain contexts such as the United Kingdom. The Education Endowment Foundation in the UK regularly commissions practice reviews on a wide range of topics. These practice reviews, as the name suggests, focus on practice-related information such as descriptions of pedagogical practices, implementation guidelines, and their impact on teachers and learners. A practice review can draw on academic and non-academic sources, as well as primary data collection, making the insights more relevant and tailored to practice. Examples of non-academic sources include professional magazines, blogs, job advertisements, policy documents, and teaching materials. For example, although not in the area of Applied Linguistics, in a recently published practice review on teacher recruitment and retention in disadvantaged areas in England (Chong, Oxley, et al., 2024), the research team synthesised information from advertisements of teaching positions to develop a typology of effective recruitment and retention strategies for schools serving socioeconomically disadvantaged areas in England. Synthesised evidence can also be easily transformed into practitioner-focused deliverables. Referring again to the Education Endowment Foundation in the UK, they have developed a teaching and learning toolkit[1] that uses evidence from meta-analyses to create resources for teachers focusing on the implementation cost, strength of evidence, and impact on learning of a wide range of pedagogical practices. Similarly, the National Institute of Teaching in England has recently beta-launched their teacher education portal, drawing on research evidence from over 300 research syntheses on teacher education literature internationally. The portal combines research evidence summaries with videos, exemplars, and practical suggestions for improving initial teacher education and teacher professional development.

Research syntheses are also useful to inform policymaking. I have recently led a rapid evidence assessment on the topic of mentoring for school leaders. This review is commissioned by the UK Government's Department for Education, who is looking for research evidence regarding various approaches to mentorship for school leaders such as principals and school trust CEOs and their impact. My team and I looked at the existing evidence and developed a framework for planning, implementing, and evaluating mentorship programmes. The synthesis report is accompanied by a proof-of-concept web-based application[2]

1. https://educationendowmentfoundation.org.uk/education-evidence/teaching-learning-toolkit

that guides schools or education providers who are interested in developing such programmes. In addition, we created an evidence and gap map website that points to areas of research that are underexplored in the international literature, informing the UK Government's funding allocation for future research on this topic.

Research syntheses are useful for practitioners and policymakers because they facilitate evidence translation and use. Not only that, research syntheses provide the opportunity for practitioners and policymakers to be involved in the synthesis process. An example is Cochrane's The People's Review project.[3] In the project, the public is invited to participate in different stages of research synthesis projects including suggesting topics for review and screening the literature. For certain types of research synthesis such as a scoping review, methodologists have suggested the involvement of stakeholders in the review process to ensure the focus and findings of the review address the needs of the stakeholders (Arksey & O'Malley, 2005).

About the book

In our editorial for a special issue on research syntheses in language education (Chong, Bond, & Chalmers, 2024), a sub-field of Applied Linguistics, we discussed the "methodological black box" of research syntheses, introducing to the field a range of research syntheses designs and methodologies, especially those that complement quantitative meta-analyses, including bibliometric and scientometric review (Zakaria & Aryadoust, 2024), meta-review (Vuogan & Li, 2024), methodological synthesis (Chalmers et al., 2024), and qualitative synthesis of research (Liu & Chong, 2024). As part of the special issue, a typology of secondary research that comprises 13 types of review is introduced (Chong & Plonsky, 2024), a subset of which is refined into three families of research synthesis in Applied Linguistics: configurative, exploratory, and explanatory reviews (Chong, 2025).

I have witnessed an increasing variety of research syntheses published in recent years in Applied Linguistics, which gives rise to the need for an open forum to discuss methodological and ethical challenges related to conducting synthetic research and ways to address them. There are emerging platforms that enable researchers to share innovative research synthesis methodologies and methodological dilemmas, including the journals *Research Synthesis in Applied Linguistics* and *Research Methods in Applied Linguistics*. At the same time, I have experienced

2. https://mentoring-for-school-leaders-app.netlify.app/

3. https://www.thepeoplesreview.ie/

some resistance to such scholarly discussions. Some see little value in reflecting on research synthesis methodologies because of the existence of well-established methodological guidelines for research syntheses including PRISMA and AMSTAR 2, as well as tools for creating search strings, inclusion and exclusion criteria, and extraction items such as PICO and SPIDER. These guidelines and tools, albeit helpful and widely used, were developed for research outside of Applied Linguistics. It does not take a genius to understand that research on drugs and on language use takes drastically different approaches, rendering some aspects of these tools less relevant to Applied Linguistics research. In Applied Linguistics, reactions to these guidelines and tools vary. Some embrace them, following strictly the steps outlined in these documents (or at least claiming to be so). Others refer to methodological guides developed specifically for Applied Linguistics research, including journal guidelines. This book proposes a third response: methodological agility in research synthesis that puts research ethics at the centre (see Chapter 19 for a framework). This book acknowledges that tensions and limitations exist in different stages of a research synthesis. While existing guidelines and tools provide a viable pathway and a tried-and-tested perspective, these challenges remain dependent on the topic, resources, time, experience of the research team, and impact plan. This book offers an insider's perspective from those who have conducted, commissioned, edited, and peer-reviewed research syntheses; contributors openly reflect on their experience, focusing on specific academic journals (Chapters 2, 3, and 4) and types of research synthesis namely bibliometric/scientometric review (Section 3), meta-analysis (Section 4), qualitative research synthesis and practice review (Section 5), scoping review (Section 6), and systematic review (Section 7). Section 8 reports emergent and interdisciplinary examples of research synthesis. Specifically, reflections and lived experiences are recorded in a structured manner for Sections 3 to 7 (and Chapter 16 in Section 8):

- **Background:** Introduction of the research synthesis project, including details of any publications
- **Methodological and ethical dilemmas/issues:** Details of the methodological and ethical questions raised, including those anticipated and those that arose in the research process
- **Course of action:** Response to the challenges by the researchers
- **Alternative courses of action:** Alternative responses and their implications for the synthesis i.e., what could have been done differently
- **Conclusion:** A summary of the lessons learned
- **Questions for discussion:** Questions that guide readers to reflect on the authors' experience and their own synthesis plans

– **Further reading:** Examples of research syntheses or methodological texts

Sections 3 to 7, each focuses on one type of research synthesis, begin with a short section overview, a "how-to" guide for conducting the specific type of research synthesis. While the focus of this book is not on how to conduct a research synthesis (as there are plenty of resources already), readers may find it helpful to have some introductory walkthroughs as an added feature of the book. These section overviews include a concise definition of the research synthesis type, an outline of key methodological stages, software or resources that can be used, and annotated examples of relevant publications.

The message of this book is to promote ethical methodological agility in research syntheses in Applied Linguistics, an idea I elaborate on in Chapter 19, driven by an attention to ethical use of approaches, methods, resources, and tools. This hybridity includes inviting methodological dialogues with synthesists in other disciplines (Chapter 18) to contemplate humbly and critically the lessons Applied Linguistics can learn from disciplines where research syntheses are more established. At the heart of this ethical methodological agility is the notion of ethics, first introduced in the seminal book by Norris and Ortega (2006), which John Benjamins also published. Norris and Ortega (2006) called for a "synthetic research ethic", which is a plea "for all of us involved in research to think and act synthetically rather than idiosyncratically… in how we go about all aspects of research, from primary to secondary, and from the perspective of knowledge generation as well as knowledge utilisation" (p. 4). While Norris and Ortega's notion of "ethics" focuses on encouraging researchers to build on one another's work and be mindful of and deliberate about the impact that their work has on society, the ethical lens that I focus on in this book concerns the meaning of research ethics in the context of research syntheses and how it can serve as a guide for researchers to make methodological decisions.

Publishing this book in 2026 is special on different levels. The year 2026 marks 20 years since research syntheses were formally introduced to Applied Linguistics by Norris and Ortega (2006). Having this book published by John Benjamins, the same publisher of "Synthesizing Research on Language Learning and Teaching" (2006), is extra special. The year 2026 is also the 50th anniversary of "meta-analysis", a type of research synthesis first introduced by Gene Glass (Glass, 1976) and 40 years after the term "best-evidence synthesis" was coined by Robert Slavin (Slavin, 1986).

AI use declaration statement

This chapter was written entirely by the author. After the draft was completed, Claude Sonnet 4.6 (Anthropic) was used for copyediting. Specifically, it was used to check for internal inconsistencies and contradictions. All observations generated by the AI tool were reviewed and verified by the author before any changes were made. The author takes full responsibility for the content of this chapter.

References

Alferink, I., & Marsden, E. (2023). OASIS: One resource to widen the reach of research in language studies. *Innovation in Language Learning and Teaching*, *17*(5), 946–952.

Arksey, H., & O'Malley, L. (2005). Scoping studies: Towards a methodological framework. *International Journal of Social Research Methodology*, *8*(1), 19–32.

Block, D. (2003). *The social turn in second language acquisition*. Edinburgh University Press.

British Educational Research Association. (2024). *Ethical guidelines for educational research* (5th ed.). British Educational Research Association.

Byrnes, H. (2013). Notes from the editor. *The Modern Language Journal*, *97*(4), 825–827.

Canagarajah, S. (2013). *Translingual practice: Global Englishes and cosmopolitan relations*. Routledge. https://www.routledge.com/Translingual-Practice-Global-Englishes-and-Cosmopolitan-Relations/Canagarajah/p/book/9780415684002

Chalmers, H., Brown, J., & Koryakina, A. (2024). Topics, publication patterns, and reporting quality in systematic reviews in language education. Lessons from the international database of education systematic reviews (IDESR). *Applied Linguistics Review*, *15*(4), 1645–1669.

Chong, S.W. (2025). Synthesis Methods and Reporting Tool (SMART) for research syntheses in Applied Linguistics. *Research Synthesis in Applied Linguistics*, *1*(1), 17–38.

Chong, S.W. (2020). The role of research synthesis in facilitating research-pedagogy dialogue. *ELT Journal*, *74*(4), 484–487.

Chong, S.W., Nie, D., & Liu, Q. (2025). Protocol for a meta-review on Applied Linguistics researchscape. *Research Synthesis in Applied Linguistics*, *1*(2), 459–469.

Chong, S. W., Bond, M., & Chalmers, H. (2024). Opening the methodological black box of research synthesis in language education: Where are we now and where are we heading? *Applied Linguistics Review*, *15*(4), 1557–1568.

Chong, S.W., Oxley, E., Negrea, V., Bond, M., Liu, Q., & Kong, M.S. (2024). *Teacher recruitment and retention in schools in socio-economically disadvantaged areas in England: Review of practice*. Education Endowment Foundation. https://d2tic4wvo1iusb.cloudfront.net/production/documents/pages/practice_review_-_recruitment_and_retention_in_disadvantaged_schools.pdf?v=1760091494

Chong, S.W., & Plonsky, L. (2024). A typology of secondary research in Applied Linguistics. *Applied Linguistics Review*, *15*(4), 1569–1594.

Chong, S.W., & Reinders, H. (2021). A methodological review of qualitative research syntheses in CALL: The state-of-the-art. *System*, *103*, 102646.

Consoli, S., & Ganassin, S. (2025). Reflexivity as a means to address researcher vulnerabilities. *Applied Linguistics Review*, *16*(6), 2521–2544.

Cushing, I. (2024). Transformative justice as a method in Applied Linguistics. *Annual Review of Applied Linguistics*, *44*, 108–117.

De Costa, P. I., & Isbell, D. R. (Eds.). (2024). *Ethics in Applied Linguistics research methods: Evidence and action* [Special issue]. *Research Methods in Applied Linguistics*, *3*(1). https://www.sciencedirect.com/special-issue/100PB0SVPZL

De Costa, P. I., Rabie-Ahmed, A., & Cinaglia, C. (2024). *Ethical issues in applied linguistics scholarship*. John Benjamins. https://benjamins.com/catalog/rmal.7?srsltid=AfmBOoovce5bVsiugsAPsfKR6kKipdMte24jhybLDtD-on6TuPxaXByf

De Costa, P. I. (2015). *Ethics in Applied Linguistics research: Language researcher narratives*. Routledge.

Dörnyei, Z., & Ryan, S. (2015). *The psychology of the language learner revisited*. Routledge.

Etymonline. (2025a). *Synthesize*. https://www.etymonline.com/word/synthesize

Etymonline. (2025b). *Analyze*. https://www.etymonline.com/word/analyze

García, O., & Li, W. (2014). *Translanguaging: Language, bilingualism and education*. Palgrave Macmillan.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher*, *5*(10), 3–8.

Gurzynski-Weiss, L., & IATBLT (n.d.). *The TBLT Language Learning Task Bank*. https://tblt.indiana.edu

Isaacs, T., & Chalmers, H. (2025). Reducing 'avoidable research waste' in Applied Linguistics research: Insights from healthcare research. *Language Teaching*, *58*(2), 214–231.

Landerholm, A. (2025). AI in qualitative health research appraisal: Comparative study. *JMIR Formative Research*, *9*, e72815.

Lantolf, J. P., & Pavlenko, A. (1995). Sociocultural theory and second language acquisition. *Annual Review of Applied Linguistics*, *15*, 108–124.

Littell, J. H. (2018). Conceptual and practical classification of research reviews and other evidence synthesis products. *Campbell Systematic Reviews*, *14*, 1–21.

Liu, M., Chong, S. W., Marsden, E., McManus, K., Morgan-Short, K., Al-Hoorie, A. H., … Hui, B. (2023). Open scholarship in Applied Linguistics: What, why, and how. *Language Teaching*, *56*(3), 432–437.

Liu, Q., & Chong, S. W. (2024). Bilingual education in China: A qualitative synthesis of research on models and perceptions. *Applied Linguistics Review*, *15*(4), 1671–1693.

MacIntyre, P. D., Gregersen, T., & Mercer, S. (Eds.). (2016). *Positive psychology in SLA*. Multilingual Matters. https://www.multilingual-matters.com/page/detail/positive-psychology-in-sla/?SF1=work_id&ST1=CVIEW-959

Norris, J., & Ortega, L. (2006). *Synthesizing research on language learning and teaching*. John Benjamins.

Plonsky, L. (Ed.). (2024a). *Open science in Applied Linguistics*. Applied Linguistics Press.

Plonsky, L. (2024b). Study quality as an intellectual and ethical imperative: A proposed framework. *Annual Review of Applied Linguistics*, *44*, 4–18.

Ruan, M., Fan, J., Liu, M., Meng, Z., Zhang, X., & Zhang, C. (2025). Artificial intelligence for the science of evidence synthesis: How good are AI-powered tools for automatic literature screening? *BMC Medical Research Methodology*, *25*(1), 199.

Sato, M., Chong, S.W., Aktar, T., Cowell, J., Kong, M.S., & Shaahdadi, M. (2024). Creating and sustaining a platform for researchers and teachers to communicate: An example of *TESOLgraphics*. *Innovation in Language Learning and Teaching*, 1–11.

Shepherd, J. (2014). *How to achieve more effective services: The evidence ecosystem*. What Works Network & Cardiff University. https://orca.cardiff.ac.uk/id/eprint/69077/1/2014_JPS_What_Works.pdf

Slavin, R.E. (1986). Best-evidence synthesis: An alternative to meta-analytic and traditional reviews. *Educational Researcher*, *15*(9), 5–11.

Ushioda, E. (2016). Language learning motivation through a small lens: A research agenda. *Language Teaching*, *49*(4), 564–577.

Vuogan, A., & Li, S. (2024). A systematic review of meta-analyses in second language research: current practices, issues, and recommendations. *Applied Linguistics Review*, *15*(4), 1621–1644.

Zakaria, A., & Aryadoust, V. (2024). A scientometric analysis of Applied Linguistics research (1970–2022): Methodology and future directions. *Applied Linguistics Review*, *15*(4), 1595–1619.

SECTION 2

Research synthesis in Applied Linguistics journals

CHAPTER 2

Research syntheses in the international journal, *Language Testing*

Publishing trends and a drive to improve research reporting standards

Talia Isaacs[1], Ruslan Suvorov[2], Xun Yan[3] & Sofya Styrina[3]
[1] University College London | [2] University of Western Ontario |
[3] University of Illinois at Urbana-Champaign

Against the backdrop of growth of research syntheses in Applied Linguistics, the international journal, *Language Testing*, introduced new article submission categories for "Systematic reviews" and "Meta-analyses" in 2022 (Harding & Winke, 2022). In 2024, in an effort to more clearly define research syntheses, increase the consistency of their reporting, and improve reporting standards, *Language Testing* introduced dedicated author guidelines for the "Systematic reviews" article type (Sage Publications, 2025). We consider "systematic reviews" to be a broad category encompassing meta-analyses, bibliometric analyses, scoping reviews, and other forms of research syntheses that include a systematic search strategy, screening, and selection process. In this chapter, we first describe how and why *Language Testing* encouraged the submission of more research syntheses and describe the development of expanded author guidelines, underpinned by a desire to catalyze improvements in reporting quality. After outlining the expectations, and requirements of the author-facing guidelines and growth trajectory of the research syntheses in the journal, we conclude by discussing next steps.

Keywords: evidence synthesis, language assessment, language testing, meta-analysis, reporting quality, secondary research, systematic reviews

https://doi.org/10.1075/rmal.17.02isa
Available under the CC BY-NC-ND 4.0 license.© 2026 John Benjamins Publishing Company

Introduction: Launch of dedicated submission categories for syntheses in *Language Testing* (2022)

The COVID-19 pandemic was a challenging time for academic journals (Suart et al., 2022), but some shifts and innovations occurred during that period that catalyzed changes in some publishing practices (e.g., Nane et al., 2023). In the case of the international, high-calibre, specialized Applied Linguistics journal, *Language Testing*, which has contributed to the professionalization of and been at the forefront of innovation in the field of language testing since its inception in 1984 (Winke, 2023), the timing of a milestone during the pandemic may actually have been serendipitous. In the midst of limited freedom of movement and restricted social interaction in 2021, a ray of hope emerged from the doldrums of the pandemic at the journal's online synchronous Editorial Board meeting, which led to changes reflected in the journal's 2022 Editorial (Harding & Winke). As online subscriptions to the journal were becoming more prominent than print subscriptions and with that trend likely to continue, journal editors were no longer bound by a fixed number of pages for each issue. Therefore, the conundrum at the 2021 Editorial Board meeting was what the Editors should do with the benefit of a flexible page budget, which effectively meant that the journal could publish more papers and/or lengthier content in each issue. At that time, the sole article types for the journal were original articles (i.e., standard-length submissions), book reviews, and test reviews. The journal description at that time did not specifically mention that it invited research syntheses or replication studies, although there was a tradition of emphasizing that it welcomed submissions of primary research studies (quantitative, qualitative, mixed methods) and theoretical papers in recent iterations.

The key recommendations of relevance to this chapter that arose at that Editorial Board meeting, which led to concrete action, was that research syntheses, which, at the time, were rarely submitted to and published in *Language Testing* (see below), seemed to be experiencing an uptick of activity in the field of Applied Linguistics more generally (later documented in e.g., Chong et al., 2024; Liu & Hu, 2024). Spurred by the first author, who notably was an Editorial Board member and not journal Editor at the time, some Editorial Board members argued for the strategic importance of using the expanded wordcount for a dedicated submission category for research syntheses so that potentially longer reference lists for included studies could be accommodated. They hoped that this would encourage an uptick of submissions of research syntheses to the journal and, thus, drive change in the field. More specifically, consolidating knowledge through research syntheses was argued to be important for a maturing field as *Language Testing*'s fortieth year anniversary was approaching (the journal's anniversary issue was

published in 2023). This recommendation received a positive sounding at the meeting and was rapidly enacted by the erstwhile Editors, with systematic reviews and meta-analyses rolled out as distinct new article types at the same time as four other shorter article types were also introduced (Harding & Winke, 2022). This milestone for the journal expanded the scope of possible article types in the journal, although some submission categories are, as yet, underutilized, including the Registered Reports category for primary and secondary research (Isaacs & Winke, 2024).

At the time of launching these new article types until the latter part of 2024, when a new set of Editors for the journal made a subsequent update, the guidance for submissions of systematic reviews and meta-analyses was highly limited and consisted of a less than 85-word blurb for each. Both specified a 12,000-word limit (as compared to 9,000 words for a standard-length article) and the journal's rationale for introducing these submission categories. For systematic reviews, the blurb emphasized that the expanded word count could "allow for sufficient scope and depth" for a high-quality review than had hitherto been possible when systematic reviews were submitted as regular standard-length articles (Harding & Winke, 2022, p. 4). While it is not possible to accurately track submission statistics of regular articles that were research syntheses due to limited search functions and archiving, we would note that no systematic review submissions had met the publication threshold by the time the 2022 Editorial was published (Harding & Winke, 2022). The first author of this book chapter had been invited to review at least two such papers during the pandemic and before the launch of the new article types, leading that Editorial Board Member and perhaps also the Editors to reflect on their quality in light of wordcount constraints for the regular submission type as an impetus for change. For meta-analyses, the blurb encouraged authors to choose *Language Testing* as the venue of choice for their research syntheses, noting that the journal had "already begun publishing" meta-analyses as regular submissions (see below section on publication trends). Neither blurb, however, defined the terms, clarified their distinguishing features, nor guided authors on the mechanics of a high-quality submission. In the next section, we detail the journal's approach to promoting higher quality reporting of research syntheses through subsequent guidelines and requirements for authors and the early implementation of those guidelines. Finally, we chart publication trends for research syntheses in *Language Testing* over time and discuss next steps and future directions.

Before progressing this chapter further, however, we add a brief terminological note. In this chapter, we use the term "research synthesis" to maintain coherence and consistency with other contributions to this volume. However, our preferred term is "evidence synthesis," which is common in science and medi-

cine, for one main reason. "Evidence synthesis" ipso facto emphasizes that it is the *evidence* that is being synthesized and consolidated. However, there are different forms of evidence that may not conform to traditional "research," which is, nonetheless, the most dominant and prominent evidence source. Some articles in medicine, for example, have interrogated what constitutes evidence (e.g., Pearson et al., 2007) and have concluded that evidence is far more diverse than the gold standard research produced in that field. Put into practice, some syntheses, including rapid reviews of evidence during the COVID-19 pandemic, may include publication types that are not considered formal research, such as editorials, websites, grey literature (e.g., non-refereed internal reports or information brochures), stakeholder consultations, expert views, or policy documents (e.g., Treweek et al., 2024). In the case of language assessment-specific articles, language tests and guidance for test users could potentially be included as publication types in synthesis searches as could datasets or corpora. From this perspective, the term "evidence synthesis" is more inclusive than "research synthesis" and, therefore, is *Language Testing*'s preferred term. We would note that our use of this term broadly aligns with Chong et al.'s (2024) description of "research synthesis," with the main difference being the semantic emphasis on evidence in its framing. We would hasten to add that all research syntheses published in *Language Testing* to date have been of primary research studies (albeit potentially coding for things like the availability of open access resources to accompany the main articles; Liu et al., 2024). However, it would be unwise to preclude drawing on evidence sources that are not traditional primary or secondary research in future syntheses, suggesting that the term "evidence synthesis" is the more appropriate term in view of future potential contributions to *Language Testing*. Through this logic, we acknowledge the agency that our editorial team has in driving top-down change in our field, and we elaborate on this point later in the chapter.

Pushing the agenda forward: Detailed guidance for authors on journal reporting requirements

By January 2024, the "change of the guard" (Isaacs & Winke, 2024, p. 3) was completed and a new Editorial team, namely, the authors of this chapter, had replaced the Editors and their associates that had initiated the new article submission categories just two years earlier (see Isaacs & Yan, 2025). While we, the new cast of Editors, applauded the introduction of the new article types and were buoyed by inheriting this development, which has contributed to the diversity of contributions published in *Language Testing*, we also recognized that the new article types were rapidly introduced during the pandemic and needed revisiting. We

were also able to observe how they were working in practice with submission/publication figures and through drawing on experience of in-house review, author queries, and so forth. On balance, we felt that some article types would benefit from being actively promoted so they would lead to more submissions (e.g., registered reports; Isaacs & Winke, 2024; and we are still hoping that they will realize their potential). For other submission categories, we recognized the pressing need to clarify our expectations about scope and quality to authors. It was clear from some submissions that some authors did not know what constituted a systematic review. Some did not understand that this term is not to be bandied about as a way of conveying rigor of a state-of-the-art review as "systematic" implies, but, rather, is a technical term for a submission category that must include some essential elements. However, the journal at the time provided no definitions nor guidance about what constituted a systematic review or meta-analysis. Thus, we recognized the pressing need to address this in an expanded set of guidelines. We were also keen to improve reporting consistency and to improve reporting standards for research syntheses published in the journal and, thereby, pave the way for higher quality submissions in the Applied Linguistics field more generally as one of the few substantively focused Applied Linguistics journals, together with *Language Learning*, that had a dedicated submission category for research syntheses (Isaacs & Chalmers, 2025).

In light of these pressing needs and our vision for improving research rigour and reporting quality for secondary research, we were the first Applied Linguistics journal, to our knowledge, to provide authors with extensive guidelines for research syntheses. Our first change was to collapse "Systematic Reviews" and "Meta-analyses" into a single submission category. Part of the logic was: Would other types of reviews, that were, as yet, less prominent in the field and not mentioned as their own submission category, also require their own submission category (e.g., methodological reviews, bibliometric)? We felt that this would be illogical and would lead to too much fragmentation. We therefore collapsed these two submission categories and opted for a single category for research syntheses, regardless of whether they were qualitative or quantitative syntheses or both. To do so, we adopted a broad or expanded definition of "systematic reviews," in line with authoritative sources (e.g., Gough et al., 2017; Higgins et al., 2024), to also include other types of research syntheses (e.g., meta-analyses, scoping reviews, bibliometric reviews, methodological syntheses, etc.). Accordingly, the "Meta-Analysis" article type on the journal's submission and peer review management system, ScholarOne, was retired, again with meta-analyses and any other forms of research syntheses to be submitted through the "Systematic review" category. We explicitly defined the scope of "systematic reviews" at the outset of the expanded guidelines as follows: "encompasses all secondary research that uses formal, rig-

orous methods to analyze data from primary (or, if many reviews already exist, secondary) research studies to consolidate and advance knowledge and understanding of a specific topic in language testing." We listed potential research synthesis types that could fall under this submission category (list non-exhaustive) then clarified the essential elements for all forthcoming submissions: "All systematic reviews must explicitly state the rationale and purpose of review, the research questions, and the inclusion and exclusion criteria used for selecting studies" (Sage, n.d.).

To inform the drafting of these expanded guidelines to mitigate the risk of having the Editors impose a set of expectations top-down that bore no relation to current submissions to the journal, we collated methodologically-oriented reviewers' comments from previous research syntheses submitted to *Language Testing*, particularly for He et al.'s (2025) scoping review. We then sought input on our draft guidelines from external experts and finalized the guidelines in light of the feedback. In addition to emphasizing principles of replicability, transparency, and rigour, including through engaging in open science practices (e.g., sharing the search strategy, extracted data, etc.; Sage, n.d.), there were some highlights of the guidelines that we particularly want to emphasize. First, we require submissions to "systematic reviews" to include a PRISMA flow diagram (Page et al., 2021), which both methodological reviewers of He et al.'s (2025) draft manuscript had recommended. This aligned with the view that it was important for readers of the article to be able to view, in one place, the number of publications identified in the search, screened, and ultimately included while documenting, for example, reasons for exclusions to improve the quality of reporting and transparency. It was difficult for readers to retrieve such information from earlier reviews published in Applied Linguistics journals (e.g., Lee et al., 2015). Therefore, being able to accurately extract the relevant information from a single figure in the form of a standardized flowchart (e.g., number of records identified, screened, and included alongside reasons for exclusions) was a notable improvement in research reporting. We would note that such information is also essential for umbrella reviews — that is, research syntheses of existing research syntheses, the next step in a high-level summary and consolidation of existing evidence on an already well-researched and well-synthesized topic.

We also require an appraisal or evaluation of the methodological quality of included studies for some forms of research synthesis because their quality is fundamentally determined by and can only be as good as the quality of the primary studies they synthesize (Gough et al., 2017). We have recently started implementing this requirement, although we recognize that quality appraisals of included studies are not appropriate for some types of research syntheses (e.g., bibliometric) and, therefore, are open to hearing from authors if evaluating method-

ological quality is unintuitive for their paper. He et al.'s (2025) scoping review noted that appraisals of methodological quality are often not required in scoping reviews, citing Peters et al. (2020). Hence, they did not include one. Chen et al. (2025) provided a bespoke appraisal of methodological quality in their systematic review of differential item functioning (DIF) by reporting the proportion of studies that had indicated whether key statistical assumptions were met and whether essential statistical indices were included. Rather than applying an established checklist, their approach focused on documenting the extent to which primary studies reported key analytical requirements. In a research synthesis on rater-mediated rubrics for interpreting, Han et al. (2026) developed a bespoke tool to evaluate methodological quality after determining that the closest available instrument, the COSMIN Risk of Bias Checklist (Mokkink et al., 2018), was designed for assessing outcome measures in clinical contexts and did not align with the aims of their study. Their custom-made tool is available as supplementary material on the Open Science Framework (OSF; Han et al., 2025). Conversely, Wang and Min (2026) made use of the COSMIN Checklist (Mokkink et al., 2018), although "the relevant indicators (e.g., correlation, kappa, and agreement) were already the core outcomes of our coding framework" (p. 176) and so overlapped with reliability indices already being systematically reported. They also investigated potential publication bias for included studies using funnel plots. Finally, Naismith and Cardwell (in press) adapted Gorard's Sieve (Gorard, 2024), a well established instrument for judging methodological quality, to enhance its applicability and relevance to their research synthesis on test component ordering (i.e., sequencing test sections). Although these studies offer different solutions to the challenge of assessing methodological quality, they collectively stand at the frontier of rigour in language testing and Applied Linguistics, where such practices are not yet widespread. They set precedents for subsequent research syntheses in the journal and beyond, demonstrating that even in the absence of a fieldwide tool, incorporating an assessment of methodological quality can enhance the rigour, transparency, and ultimately the trustworthiness of a research synthesis.

Another element of the journal guidelines (Sage, n.d.), which is currently a recommendation rather than a requirement, is that authors pre-register their systematic reviews. Doing so is essential to safeguard against publication bias (Chalmers et al., 2024). Pre-registration could happen, for example, through the International Database of Education Systematic Reviews (IDESR, n.d.) for topics relevant to education (Chalmers et al., 2024), on PROSPERO for medically oriented papers, on OSF, or as a Registered Report in *Language Testing* or another peer reviewed journal (e.g., *Systematic Reviews, Research Synthesis in Applied Linguistics*, etc.; note that this article type is called a Research Protocol in some journals). However, we have not yet implemented the pre-registration recommendation, mainly

because reviews that reach us are already at full dissemination phase and do not represent a research synthesis in progress, which is when pre-registration is appropriate. Journal editors in Applied Linguistics need encourage a culture of pre-registration, whether for primary or secondary research. They may incentivize it with an in-principle accept of the full-length paper when a registered report is published (Isaacs & Winke, 2024) and find different channels to promote it as an article type, although this has not yet worked for *Language Testing*.

Another point that we require but have not yet been able to enforce is that *Language Testing* authors who publish on topics to do with education use IDESR to check whether previous related research syntheses have been conducted when conceptualizing and reporting on their research synthesis. A review centring on a healthcare topic would need to do the same using PROSPERO, although we have not yet referred to this in our guidelines, which we will evolve as new developments and capabilities take shape and in response to research syntheses that break the mold. Being aware of and building on primary and secondary research that has been published previously is criterial for high quality scholarship and essential for preventing the unnecessary duplication of efforts, which is a form of "research waste" in Applied Linguistics (Isaacs & Chalmers, 2025; Plonsky et al., 2025) and other fields (Glasziou & Chalmers, 2018; Purgar et al., 2022).

In the next section, we overview publication trends for research syntheses in the journal since its inception in 1984 until present (2025).

Trends in the publication of research syntheses in *Language Testing* (1984–2025)

Search strategy, title/abstract screening, and the limitations of our approach

To take stock of publishing trends of research syntheses published in *Language Testing* from 1984 to 2024, we ran a search in December 2024 using the Educational Resources Information Center (ERIC) research repository using the following search terms and Boolean operators: "systematic review" OR "meta-analy*" OR "scoping" OR "bibliometric" OR "synthesis." This search led us to identify 18 studies that we extracted to a spreadsheet and screened further. Four entries that were pulled up in the initial search were excluded from being classified as research syntheses and, therefore, do not feature in Table 1 and Figure 1 below. One of these (Saito, 2008) provided a brief (300-word plus table) meta-analytic summary of four studies within the literature review, reporting inclusion criteria and effect sizes. However, the primary focus of the manuscript was on two empirical studies, examining the effect of training on peer assessment of second

language (L2) oral presentations. Therefore, synthetic methods were embedded within the literature review, but the main focus and overall orientation of the publication was on reporting primary research studies. The other studies did not conform with the definition of "systematic reviews" in our revised guidelines (Sage, n.d.). One cited a systematic review in the abstract to highlight a research gap but was itself a primary research study (Nicklin & Vitta, 2022). Another that used the term "discourse synthesis" (Plakans, 2009, p. 561) was also a primary research study. The fourth study reported on a "systematic accessibility review process" (Christensen et al., 2023, p. 856), characterizing the authors' approach to developing an accessibility checklist for a kindergarten English language proficiency test, but again lacked research synthesis elements (Chong et al., 2024).

We acknowledge the limitations associated with our search, conducted expressly for the purpose of this chapter. It is possible that our search strategy may not have picked up papers that were not labelled as research syntheses in titles or abstracts. In addition, we only screened titles and abstracts and did not evaluate full-text articles. Therefore, we cannot discount the possibility of some inadvertent omissions. Nonetheless, the search provided us with indicative trends of research synthesis publications in the journal over time, which was sufficient for the descriptive purpose of our chapter.

Once the initial search and screening were completed, we manually included research syntheses that had been published in all 2025 issues and the first two (of four) issues in 2026 to ensure that we presented the most up to date figures available at the time of writing. This resulted in the inclusion of four additional entries. Note that the date of publication for all outputs was determined by when the article was assigned to a print issue and was published in its final form and not by when it initially appeared as an advance online publication.

Published research syntheses in "Language Testing": Taking stock

Based on our search, since its inception, *Language Testing* has published a total of 16 research syntheses to date (end of March 2026). The first, a meta-analysis of research on self-assessment in L2 testing by Steven Ross, appeared in 1998. While published research syntheses remained relatively rare in the journal for its first three decades, there has since been a notable overall increase, with 8 out of 14 research syntheses appearing in print issues from 2021 to 2025 and the journal on course to match or exceed its highest annual figure of three in 2026. This mirrors the growth trend in submissions of research syntheses in Applied Linguistics more generally (e.g., Liu et al., 2024), though in smaller numbers, and may also reflect the growing recognition of their value for synthesizing evidence and

consolidating knowledge in Applied Linguistics research, including in language assessment (see Plonsky et al., 2025).

Figure 1 charts the frequency of research syntheses published in print in *Language Testing* by year from the inception of the journal in 1984 through to the end of 2025. Aside from Ross' (1998) early meta-analysis and a second meta-analysis published over a decade later (In'nami & Koizumi, 2009), there were no publications of research syntheses that appeared in the first three decades of journal publication.

The year 2014 seems to mark the start of an uptick. Notably, the introduction of the new article types for research syntheses reported in the 2022 Editorial (Harding & Winke) did not have an immediate effect on their publication that year after a peak of three in 2021. Naturally, it takes time for new journal initiatives to have an impact in light of the lengthy peer review to publication cycle. In addition, Figure 1 takes into account publication of research syntheses in print issues rather than their appearance OnlineFirst, so any effects of the change can be seen later. Finally, unusual pandemic-related circumstances may also have affected submissions and peer review during that period (e.g., Suart et al., 2022). Notably, *Language Testing*'s expanded guidance on publishing "systematic reviews," which were introduced in the latter part of 2024, were not intended to increase the volumes of submissions as the 2022 article type expansion was. Rather, that intervention was designed to contribute positively to the *quality* of the published research syntheses.

The 16 research syntheses that have been published to date (March 2026) encompass a range of review types (see Table 1), with the most common being meta-analysis ($n = 8$), including most recently a research synthesis mentioning both systematic review and meta-analysis in the title following reviewer confusion about both being subsumbed under the term "systematic review" (Han et al., 2026) and "a three-level meta-analysis" (Wang & Min, 2026). Qualitative systematic reviews with no meta-analytic component also feature prominently, with five studies falling into this category, including one described as a "systematic examination" (Liu et al., 2024). The remaining review types comprise "synthesis" (In'nami & Koizumi, 2016), "comprehensive review" (Aryadoust et al., 2021), "scoping review" (He et al., 2025), and "narrative review" (Yan et al., 2016, which also included a meta-analysis).

Recent years have also seen advances in research transparency and reporting standards both seemingly linked to and independent of the publication of *Language Testing*'s "Guidelines for systematic reviews" (Sage, n.d.). In particular, the six most recent research syntheses published in *Language Testing* (Chen et al., 2025; Han et al., 2026; He et al., 2025; Ihlenfeldt & Rios, 2023; Liu et al., 2024; Wang & Min, 2026) all mentioned utilizing the PRISMA guidelines (Page et al.,

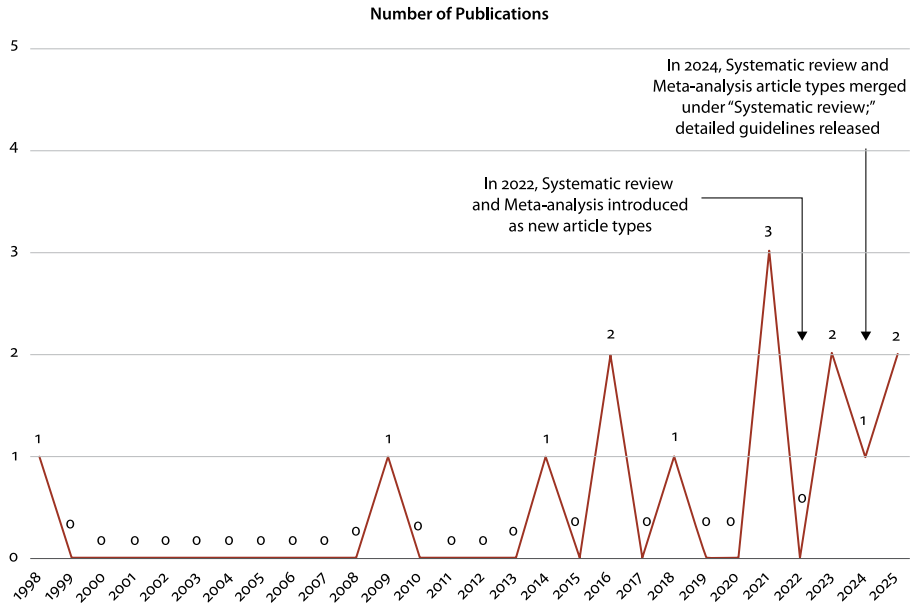


Figure 1. Number of research syntheses published in the journal, *Language Testing* (1984–2025)

Note. The figure omits 2026 because the calendar year is not yet complete at the time of writing.

2021). With the exception of Inlenfeldt and Rios (2023), which was accepted for publication before the journal's "Systematic review guidelines" were released, all included a PRISMA flow diagram to illustrate the study selection process. In the case of both He et al. (2025) and Liu et al. (2024), methodologically-oriented reviewers recommended the incorporation of PRISMA flow diagrams during peer review. By the time that Chen et al. (2025) was submitted on ScholarOne, the Editors had formalized the inclusion of the PRISMA flow diagram by adding it as a checklist item for internal/in-house review, with a requirement to include it before the article is sent out for peer review. We would note that including the flow diagram, while an improvement on research syntheses that do not include it, does not imply full adherence to PRISMA guidelines.

Of the articles that referred to PRISMA (Page et al., 2021) in the body of the paper, three have included the 27-item PRISMA checklist in a Supplementary Appendix at the Editor's request (Chen et al., 2025; Han et al., 2026; Wang & Min, 2026). This sets a precedent for other articles to follow in adhering to established best practice recommendations in research reporting for research syntheses (see Chalmers & Isaacs, in press) by essentially showing where in the article each checklist item has been addressed. The inclusion of the PRISMA checklist in

Table 1. Research syntheses published in the journal, *Language Testing*, from June 1984 (inaugural issue) to March 2026

#	Author(s)	Year	Synthesis type based on description in published article	Synthesis type based on Chong & Plonsky's (2024) typology	Main topic
1	Ross	1998	Meta-analysis	Meta-analysis	Correlations between self-assessment and measures of L2 proficiency
2	In'nami & Koizumi	2009	Meta-analysis	Meta-analysis	Test format effects on reading and listening test performance
3	Koo et al.	2014	Meta-analysis	Meta-analysis	Differential item functioning (DIF) trends in a reading test
4	Yan et al.	2016	Narrative review, Meta-analysis	Mixed review	Elicited imitation as a measure of L2 proficiency
5	In'nami & Koizumi	2016	Synthesis	Systematic literature review	Task and rater effects in L2 speaking and writing
6	Wind & Peterson	2018	Systematic review	Systematic literature review	Methods for evaluating rating quality
7	Knoch et al.	2021	Systematic review	Systematic literature review	Rating scale development processes
8	Li & Zhang	2021	Meta-analysis	Meta-analysis	Correlation between self-assessment and language performance
9	Aryadoust et al.	2021	Comprehensive review	State-of-the-art review†	Rasch measurement in L2 assessment
10	Aryadoust & Luo	2023	Systematic review	Systematic literature review	Conceptualizations and operationalisations of L2 listening constructs
11	Ihlenfeldt & Rios	2023	Meta-analysis	Meta-analysis	L2 proficiency tests as predictors of academic success
12	Liu et al.	2024	Systematic examination	State-of-the-art review†	Open access status of manuscripts in LT and LAQ

Table 1. *(continued)*

#	Author(s)	Year	Synthesis type based on description in published article	Synthesis type based on Chong & Plonsky's (2024) typology	Main topic
13	Chen et al.	2025	Systematic review	Systematic literature review	DIF in L2 assessment
14	He et al.	2025	Scoping review	Scoping review	L2 test preparation research
15	Han et al.	2026	Meta-analysis	Meta-analysis	Scoring rubrics in interpreting assessment
16	Wang & Min	2026	Meta-analysis	Meta-analysis	Human–machine correlation of automated speech evaluation

Note.
† We considered the classification of "state-of-the-art review" based on Chong & Plonsky's (2024) typology more appropriate than systematic literature review because of the large dataset in these studies from the sources where the search was undertaken.

the guidelines are currently framed as optional, although including an appropriate reporting tool may be prescribed in amended guidelines in the future. We would hasten to add that since the time that Chen et al. (2025) was in production for initial publication OnlineFirst in 2024, the new methodological journal, *Research Synthesis in Applied Linguistics*, was founded and something of an alternative to the PRISMA checklist specifically designed for Applied Linguistics researchers was launched by one of the founding Editors through the SMART (Synthesis Methods and Reporting Tool) research instrument (Chong, 2025). Again, *Language Testing* Editors update guidelines periodically and may offer authors the option of using the PRISMA checklist, SMART, or a suitable alternative research reporting tool in the future. In this way, we Editors have a role and responsibility to play in promoting research quality and ensuring that new developments in the field are integrated into the journal's way of working to keep pace with developments and remain at the cutting edge (see also Plonsky et al., this volume).

Thematically, the 16 published research syntheses in *Language Testing* have explored a variety of topics. The earliest meta-analysis by Ross (1998) examined the correlations between self-assessment and L2 proficiency, a topic later revisited by Li and Zhang (2021). The second meta-analysis published in *Language Testing* more than a decade later (In'nami & Koizumi, 2009) focused on test formats and their impact on listening and reading test performance. The issue of fairness and bias has been addressed in studies on DIF, including Koo et al.'s (2014) meta-

analytic examination of DIF trends in a reading test and Chen et al.'s (2025) systematic review of DIF in L2 assessment. Another meta-analysis by In'nami and Koizumi (2016) explored task and rater effects on L2 speaking and writing outcomes. The rater or rating theme was also the focus of four research syntheses: Wind and Peterson (2018), who investigated methods for evaluating rating quality, Knoch et al. (2021), who reviewed research on rating scale development processes, Han et al. (2026), who synthesized research on scoring rubrics for language interpreting, and Wang and Min (2026), who focused human–machine correlations of scores in the domain of automated speech evaluation.

Several research syntheses have also addressed specific conceptual and methodological aspects of language testing. In particular, Aryadoust et al. (2021) provided a comprehensive review of Rasch measurement in L2 assessment, while Aryadoust and Luo's (2023) systematic review examined how L2 listening constructs are conceptualized and operationalized in research and practice. The predictive validity of high-stakes language proficiency tests for academic success was another aspect meta-analyzed by Ihlenfeldt and Rios (2023). More recently, the scope of research syntheses has broadened to include issues of open access and test preparation: Liu et al. (2024) systematically examined the open access status of manuscripts in *Language Testing* and *Language Assessment Quarterly*, whereas He et al.'s (2025) scoping review investigated three decades of research on practices, perceptions, and impacts of test preparation. Wang and Min's (2026) three-level meta-analysis offered an up to date quantitative account of human–machine agreement in L2 speaking assessment and Han et al. (2026) elaborated a conceptual-analytic framework for rubric-mediated interpreting assessment. Collectively, the 16 research syntheses published in *Language Testing* indicate the increasing use of secondary research methods to consolidate and critically evaluate knowledge across a diverse range of topics and issues.

Concluding remarks

In conformity with this chapter's goal of detailing the international journal, *Language Testing*'s approach to encouraging more, longer (if needed, or rather, more comprehensive), and higher quality publications of research syntheses as a means of consolidating evidence in the field of language testing, in the sections above, we have detailed the journal's approach and contribution to promoting research synthesis in language testing and Applied Linguistics and the journal's leadership and (at the time) groundbreaking initiatives in this domain. After recounting the story of *Language Testing*'s drive to that effect from the days of the pandemic until present, we demonstrate both growing trends in submissions over the past decade

compared to scant publications of research syntheses in the decades before, and the growth potential of research syntheses in terms of volume, quality, and standardization of reporting practices.

We note the limitations of our search for research syntheses published in *Language Testing* over time, which, we surmise was sufficient to view trends, albeit without being able to rule out the possibility of inadvertently excluding some research syntheses, due primarily to our title/abstract search and no full-length screening of published articles that were not pulled up in the search. Our results underscore the importance of labelling research syntheses as such and using precise and appropriate terminology when referring to research syntheses (e.g., the descriptions of different forms of research synthesis in the journal's guidelines or drawing on the categories from Chong & Plonsky's typology, 2024). We have provided some examples from our search results of how imprecise language can create ambiguity and result in the undue inclusion of irrelevant studies in this and future literature searches in addition to the possible inadvertent exclusion of studies that warrant inclusion.

We anticipate a rise in quality and professionalization trend in the future in *Language Testing* as the bar of required elements has grown following the release of our 2024 guidelines (Sage, n.d.) and will likely continue to increase. We will continue to revise guidelines on a rolling basis, taking stock of developments in Applied Linguistics and, crucially, our experience of working with authors and reviewers in relation to research syntheses as we continue to evolve our approach. As above, we expect the guidelines described above and higher editorial standards to have a positive effect on the quality of research syntheses going forward and, in fact, are already witnessing this. Based on the number of submissions in the pipeline at the time of writing (Naismith & Ramsey, in press, currenty in production; Zhang & Aryadoust, 2026, available; other submissions currently undergoing peer review), we anticipate a continued stream of research syntheses that abide by the new guidelines and, as such, are likely to demonstrate greater research reporting rigour and the inclusion of standard elements compared to earlier such publications.

A final point is that our focus so far has been squarely on author-related guidelines. However, we acknowledge that there is a dire need for clear reviewer guidelines for research syntheses in addition to other article types. A key medium-term goal for the Editorial team is, therefore, to build the base of support and guidance documents available on our "Reviewer resources" page (Sage, 2025). Some exemplar reviews that were produced as part of a pilot of transparent reviews (Isbell & Kremmel, 2024) are available on this webpage, for example, including one research synthesis (Liu et al., 2024). We hope to populate this page with reviewer guidelines before long that are responsive to the issues that arise

during peer reviews, reflect the demands we require from authors and best practice recommendations, and that can help enhance the quality and consistency of peer review reports.

Disclosure of interests

The authors are members of the editorial team for the journal, *Language Testing*, and receive honoraria from Sage, the journal's publisher. The first author was a member of the Editorial Board of *Language Testing* prior to her appointment as Editor in 2023 and recounts her experience during this time period. She has also served as a member of the Editorial Advisory Board of *Research Synthesis in Applied Linguistics* since 2025.

Acknowledgments

We are grateful to Sin Wang Chong for the opportunity to showcase *Language Testing*'s approach to publishing research syntheses. We sincerely thank Sin Wang, Hamish Chalmers, and Vahid Aryadoust for their invaluable feedback on earlier versions of our guidelines for authors.

AI use declaration statement

The authors confirm that no AI tools were used in the preparation, writing, formatting, and editing of this work.

References

(* represents research syntheses published in *Language Testing*, from Issue 1, 1984 to Issue 2, 2026)

*Aryadoust, V., & Luo, L. (2023). The typology of second language listening constructs: A systematic review. *Language Testing*, *40*(2), 375–409.

*Aryadoust, V., Ng, L. Y., & Sayama, H. (2021). A comprehensive review of Rasch measurement in language assessment: Recommendations and guidelines for research. *Language Testing*, *38*(1), 6–40.

Chalmers, H., Brown, J., & Koryakina, A. (2024). Topics, publication patterns, and reporting quality in systematic reviews in language education. Lessons from the international database of education systematic reviews (IDESR). *Applied Linguistics Review*, *15*(4), 1645–1669.

Chalmers, H., & Isaacs, T. (in press). Improving reporting quality to reduce research waste in applied linguistics: More lessons from healthcare research. *Annual Review of Applied Linguistics*.

*Chen, X., Aryadoust, V., & Zhang, W. (2025). A systematic review of differential item functioning in second language assessment. *Language Testing*, *42*(2), 193–222.

Chong, S. W. (2025). Synthesis Methods and Reporting Tool (SMART) for research syntheses in Applied Linguistics. *Research Synthesis in Applied Linguistics*, *1*(1), 17–38.

Chong, S. W., Bond, M., & Chalmers, H. (2024). Opening the methodological black box of research synthesis in language education: Where are we now and where are we heading? *Applied Linguistics Review*, *15*(4), 1557–1568.

Chong, S. W., & Plonsky, L. (2024). A typology of secondary research in applied linguistics. *Applied Linguistics Review*, 15(4), 1569–1594.

Christensen, L. L., Shyyan, V. V., & MacMillan, F. (2023). Toward a systematic accessibility review process for English language proficiency tests for young learners. *Language Testing*, *40*(4), 856–876.

Glasziou, P., & Chalmers, I. (2018). Research waste is still a scandal — An essay by Paul Glasziou and Iain Chalmers. *BMJ*, 363, k4645.

Gorard, S. (2024). Judging the relative trustworthiness of research results: How to do it and why it matters. *Review of Education*, *12*, e3448.

Gough, D., Oliver, S., & Thomas, J. (Eds.). (2017). *An introduction to systematic reviews* (2nd ed.). Sage.

Han, C., Jiang, M., & Qionglu, C. (2025, October 7). *A systematic review and meta-analysis of rubrics in rater-mediated assessment of language interpreting.*

*Han, C., Jiang, M., & Chen, Q. (2026). Rubricizing the assessment practice: A systematic review and meta-analysis of rubrics in rater-mediated assessment of language interpreting. *Language Testing*, *43*(2), 197–232.

Harding, L., & Winke, P. (2022). Innovation and expansion in language testing for changing times. *Language Testing*, *39*(1), 3–6.

*He, S., Sénécal, A. M., Stansfield, L., & Suvorov, R. (2025). A scoping review of research on second language test preparation. *Language Testing*, *42*(1), 11–47.

Higgins, J. P. T., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M. J., & Welch, V. A. (Eds.). (2024). *Cochrane handbook for systematic reviews of interventions* (Version 6.5, updated August 2024). Cochrane. https://www.training.cochrane.org/handbook

IDESR. (n.d.). International database of education systematic reviews. https://idesr.org/

*Ihlenfeldt, S. D., & Rios, J. A. (2023). A meta-analysis on the predictive validity of English language proficiency assessments for college admissions. *Language Testing*, *40*(2), 276–299.

In'nami, Y., & Koizumi, R. (2009). A meta-analysis of test format effects on reading and listening test performance: Focus on multiple-choice and open-ended formats. *Language Testing*, *26*(2), 219–244.

In'nami, Y., & Koizumi, R. (2016). Task and rater effects in L2 speaking and writing: A synthesis of generalizability studies. *Language Testing*, *33*(3), 341–366.

Isaacs, T., & Chalmers, H. (2025). Reducing 'avoidable research waste' in Applied Linguistics research: Insights from healthcare research. *Language Teaching*, *58*(2), 214–231.

Isaacs, T., & Winke, P. M. (2024). Purposeful turns for more equitable and transparent publishing in language testing and assessment. *Language Testing*, *41*(1), 3–8.

Isaacs, T., & Yan, X. (2025). Language Testing behind the scenes: New capabilities, groundbreaking developments, and disruptive forces. *Language Testing*, *42*(1), 3–10.

Isbell, D. R., & Kremmel, B. (2024). Open Science and the next generation of language testing research. *Language Testing*, *41*(4), 898–908.

*Knoch, U., Deygers, B., & Khamboonruang, A. (2021). Revisiting rating scale development for rater-mediated language performance assessments: Modelling construct and contextual choices made by scale developers. *Language Testing*, *38*(4), 602–626.

*Koo, J., Becker, B. J., & Kim, Y. S. (2014). Examining differential item functioning trends for English language learners in a reading test: A meta-analytical approach. *Language Testing*, *31*(1), 89–109.

Lee, J., Jang, J., & Plonsky, L. (2015). The effectiveness of second language pronunciation instruction: A meta-analysis. *Applied Linguistics*, *36*(3), 345–366.

*Li, M., & Zhang, X. (2021). A meta-analysis of self-assessment and language performance in language testing and assessment. *Language Testing*, *38*(2), 189–218.

*Liu, M., Al-Hoorie, A. H., & Hiver, P. (2024). Open access in language testing and assessment: The case of two flagship journals. *Language Testing*, *41*(4), 703–728.

Liu, Y., & Hu, G. (2024). Research trends in Applied Linguistics (2017–2021): A scientometric review of 42 journals. In H. Meihami & R. Esfandiari (Eds), *A scientometrics research perspective in Applied Linguistics*. Springer.

Mokkink, L. B., de Vet, H. C. W., Prinsen, C. A. C., Patrick, D. L., Alonso, J., Bouter, L. M., & Terwee, C. B. (2018). COSMIN Risk of Bias checklist for systematic reviews of patient-reported outcome measures. *Quality of Life Research: An International Journal of quality of Life Aspects of Treatment, Care and Rehabilitation*, *27*(5), 1171–1179.

Naismith, B., & Cardweel, R. (in press). A systematic review of test component ordering with implications for language assessment. *Language Testing*.

Nane, G. F., Robinson-Garcia, N., van Schalkwyk, F., & Torres-Salinas, D. (2023). COVID-19 and the scientific publishing system: Growth, open access and scientific fields. *Scientometrics*, *128*, 345–362.

Nicklin, C., & Vitta, J. P. (2022). Assessing Rasch measurement estimation methods across R packages with yes/no vocabulary test data. *Language Testing*, *39*(4), 513–540.

Page, M. J., Moher, D., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., & McKenzie, J. E. (2021). PRISMA 2020 explanation and elaboration: Updated guidance and exemplars for reporting systematic reviews. *The BMJ*, *372*(160), 1–36.

Pearson, A., Wiechula, R., Court, A., & Lockwood, C. (2007). A re-consideration of what constitutes "evidence" in the healthcare professions. *Nursing Science Quarterly*, *20*(1), 85–88.

Peters, M. D. J., Godfrey, C., McInerney, P., Munn, Z., Tricco, A. C., & Khalil, H. (2020). Scoping reviews. In E. Aromataris & Z. Munn (Eds.), *JBI manual for evidence synthesis* (pp. 406–451). JBI.

Plakans, L. (2009). Discourse synthesis in integrated second language writing assessment. *Language Testing*, *26*(4), 561–587.

Plonsky, L., Hu, Y., & Zhao, Shujing. (2026). Research synthesis and meta-analysis in *SSLA*: Evidence, quality, and ways forward. In S.-W. Chong (Ed.), *Research synthesis methodologies in Applied Linguistics*. John Benjamins. (this volume).

Plonsky, L., Isaacs, T., & Taylor, L. (2025). Opportunities and challenges for research synthesis in applied linguistics: Three scholars reflect. *Research Synthesis in Applied Linguistics*, *1*(1), 5–16.

Purgar, M., Klanjscek, T., & Culina, A. (2022). Quantifying research waste in ecology. *Nature Ecology & Evolution*, *6*(9), 1390–1397.

*Ross, S. (1998). Self-assessment in second language testing: A meta-analysis and analysis of experiential factors. *Language Testing*, *15*(1), 1–20.

Sage. (2025). '*Language Testing*': *Reviewer resources*. https://journals.sagepub.com/page/ltj/reviewer_resources

Sage. (n.d.). Guidelines for systematic reviews. Retrieved 7 July 2025 from https://journals.sagepub.com/pb-assets/cmscontent/ltj/LT%20Guidelines%20for%20Systematic%20Reviews_October_3_24-1739525675.pdf

Saito, H. (2008). EFL classroom peer assessment: Training effects on rating and commenting. *Language Testing*, *25*(4), 553–581.

Suart, C., Neuman, K., Truant, R. (2022). The impact of the COVID-19 pandemic on perceived publication pressure among academic researchers in Canada. *PLoS ONE*, *17*(6), e0269743.

Treweek, S., Brazzelli, M., Crosse, A., Daga, S., Isaacs, T., Sunga, R., Brazzelli, M., Cruikshank, M., Inamura, M., Manson, P., Robertson, C., Treweek, S., Crosse, A., Sunga, R., Rishi, A., Rishi, K., Isaacs, T., Kwaku-Odoi, C., Bobat, M., … Cueva, S. (2024). Using the GRADE evidence to decision framework to reach recommendations together with ethnic minority community organizations: The example of COVID-19 vaccine uptake in the United Kingdom. *Journal of Clinical Epidemiology*, *168*.

*Wang, Y., & Min, S. (2026). The human–machine correlation in automated speech evaluation: A three-level meta-analysis. *Language Testing*, *43*(2), 165–196.

*Wind, S.A., & Peterson, M.E. (2018). A systematic review of methods for evaluating rating quality in language assessment. *Language Testing*, *35*(2), 161–192.

Winke, P.M. (2023). Forty years of Language Testing, and the changing paths of publishing. *Language Testing*, *40*(1), 3–7.

*Yan, X., Maeda, Y., Lv, J., & Ginther, A. (2016). Elicited imitation as a measure of second language proficiency: A narrative review and meta-analysis. *Language Testing*, *33*(4), 497–528.

Zhang, S., & Aryadoust, V. (2026). Extrinsic drivers of scientific impact in second language assessment: A bibliometric path analysis. *Language Testing*. Advance online publication.

CHAPTER 3

Research synthesis and meta-analysis in *SSLA*

Evidence, quality, and ways forward

Luke Plonsky, Yuhang Hu & Shujing Zhao
Northern Arizona University

This chapter examines the role of one journal — *Studies in Second Language Acquisition* (SSLA) — in applying and advancing synthetic evidence and methods in the field of Applied Linguistics. In particular, we describe a comprehensive sample of synthetic research published in SSLA in terms of its frequency and substantive foci. We then consider the quality of synthetic efforts in SSLA as presented in Hu's (2024) secondary synthesis of meta-analyses published in SSLA. We also highlight a number of methodological developments in synthetic research put forward in SSLA such as meta-analytic replication, interrater reliability estimates, methodological syntheses, adjusting meta-analytic effects for psychometric artifacts, multilevel and multivariate meta-analysis (Norouzian & Bui, 2024), and bibliometric analyses (Demi & Kartal, 2022). The chapter closes with specific and actionable recommendations for advancing synthetic efforts as a means to arriving at higher validity synthetic evidence for all of Applied Linguistics.

Keywords: SLA, SSLA, meta-analysis, methodological synthesis, effect size

1. Introduction

Meta-research is undoubtedly now a part of the empirical landscape of Applied Linguistics. What began as two main and overlapping types of secondary research (i.e., research synthesis and meta-analysis) (Norris & Ortega, 2000), has expanded into a wide variety of synthetic techniques and applications, ranging from substantive to methodological, theoretical to practice-oriented, and everything in between (Chong & Plonsky, 2024). To date, several hundred secondary reviews and meta-analyses have been published and presented, covering virtually every topic in the field, sometimes many times over.

Further evidence for the strong presence of synthetic research in the field is abundant. For example, several universities, including Northern Arizona Uni-

https://doi.org/10.1075/rmal.17.03plo
Available under the CC BY-NC-ND 4.0 license.© 2026 John Benjamins Publishing Company

versity and University of Maryland College Park, regularly offer semester-long seminars taught on meta-analysis. It is common to see articles discussing meta-analytic methods. And many different domains within the field have been meta-analyzed and re-meta-analyzed using meta-analytic replication (see Plonsky, 2015). In addition, the range and sophistication of synthetic methods have seen incredible advances (Plonsky et al., 2023)

One key to the advancement of synthetic research in Applied Linguistics has been the visibility of this type of work in academic journals. Some journals such as *Language Learning* have dedicated strands for secondary research. And indeed, ever since the publication of Norris and Ortega's (2000) landmark meta-analysis on the effectiveness of L2 instruction, members of the field have associated *Language Learning* with research synthesis and meta-analysis. Other journals (e.g., *Applied Linguistics*; *Applied Linguistics Review*; *Studies in Second Language Learning and Teaching*) have published special issues on meta-analysis and, in the case of *Language Testing*, have provided author guidelines for systematic reviews (see Isaacs et al., this volume). And, of course, thanks to the leadership of Sin-Wang Chong and Sathena Chan, the field now enjoys the benefits of an entire journal dedicated to secondary research and the methods thereof: *Research Synthesis in Applied Linguistics* (see Chan & Chong, this volume). In line with its mission to publish highly rigorous research that advances our understanding of second language development, *Studies in Second Language Acquisition* (SSLA) has also contributed in a significant way to the field's body of secondary/synthetic research. It is that line of empirical and methodological inquiry that this chapter seeks to explore.

The main body of this chapter presents, first, a summary of research syntheses published to date in one particular journal, *Studies in Second Language Acquisition* (SSLA). Specifically, we provide an overview of the types and topics of research synthesis that have appeared throughout the life of the journal. We then follow up on this overview in Section 2 by examining the methodological quality in a subset of those same studies. Specifically, drawing on a subset of the data from Hu (2024), we assess the quality of meta-analyses published in SSLA. The analysis and corresponding suggestions for future research focus on six major stages of conducting a meta-analysis as well as transparency-related practices.

2. Number, types, and topics of meta-research published in SSLA

A comprehensive literature search was conducted for all research synthesis articles published in SSLA since its first issue in 1979. We adopted the secondary research categories developed by Chong and Plonsky (2024) as a starting point

for our coding scheme. All included studies demonstrated systematic procedures in terms of literature search, data coding, and reporting of results.

As of August 2025, SSLA published 37 research synthesis articles, more than half of which appeared within the past decade (see Figure 1 below). Specifically, the period from 2021 to the present witnessed a twofold increase compared to 2016–2020, with 19 research syntheses published in less than five years. By contrast, the journal's earlier record of synthetic research was sparse: Between 1979 and 1990, no systematic research synthesis was published, and prior to 2020, typically only two or three research synthesis articles appeared per five-year period. This trend shows that research synthesis has become part of the mainstream disciplinary dialog in SSLA. More broadly, these findings are indicative of the maturity of second language acquisition (SLA) and Applied Linguistics more broadly, since research synthesis can only thrive in the context of a robust and well-developed body of accumulated primary research. The surge may also be a sign of increasing methodological rigor and transparency, enabling systematic syntheses that were previously not feasible due to inconsistency in research designs, incomplete reporting of results, and so forth. Meanwhile, it is important to note that some of the existing meta-analyses have relatively small sample sizes, which often reflects the limited availability of primary studies in certain subdomains. The full list of studies included in this sample can be found in the Appendix.

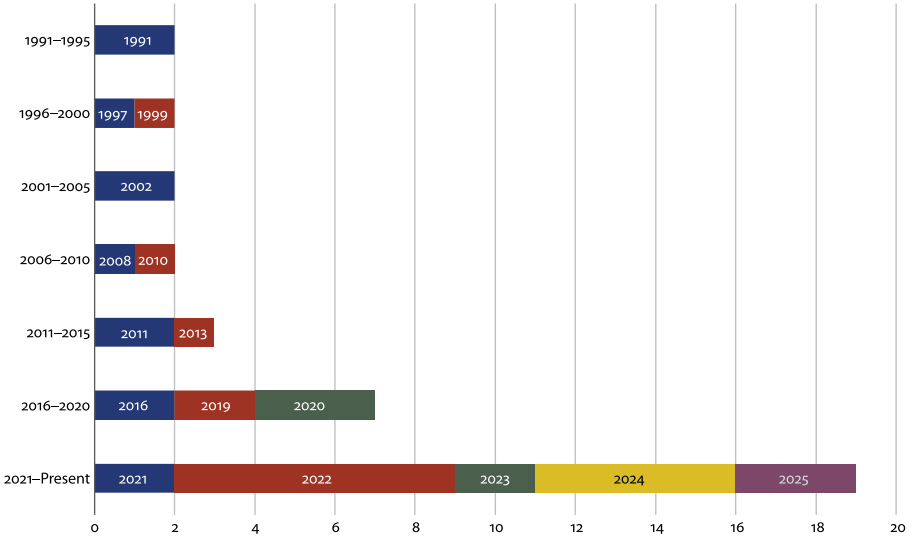


Figure 1. Number of research syntheses published by SSLA across five-year periods
Note: The final period includes only through August 2025.

In terms of research types, five categories were identified following Chong and Plonsky (2024): meta-analysis, systematic review, methodological review, bibliometric study, and scoping review. In addition, we identified and included here an additional category for articles introducing and/or discussing meta-analytic methods, which represents a unique contribution of SSLA to advancing methodological discussions in meta-analytic research. Figure 2 below shows a more detailed overview. Meta-analysis is by far the most common type, comprising nearly half of all research syntheses with 15 articles in total. The rise, particularly since 2016, indicates the expansion of primary research datasets and reflects the field's growing emphasis on quantitative synthesis and generalizability. Bibliometric studies, although only emerging recently, show a similar upward trend (see Sai & Aryadoust, this volume). This suggests a shift toward mapping the landscape of SLA, as researchers are seeking to understand not only findings but also citation practices, networks, and disciplinary development. Articles explicitly focusing on meta-analytic methods did not appear until the past five years, yet they have since become one of the major components of recently published research syntheses. Systematic reviews and methodological reviews were more prevalent in the early 2000s, serving as tools to taking stock and methodological consolidation before the surge of quantitative meta-analyses. Scoping review is the newest and least common type, with only one example to date, showing potential for future growth, especially in capturing broad or emergent topics in SLA.

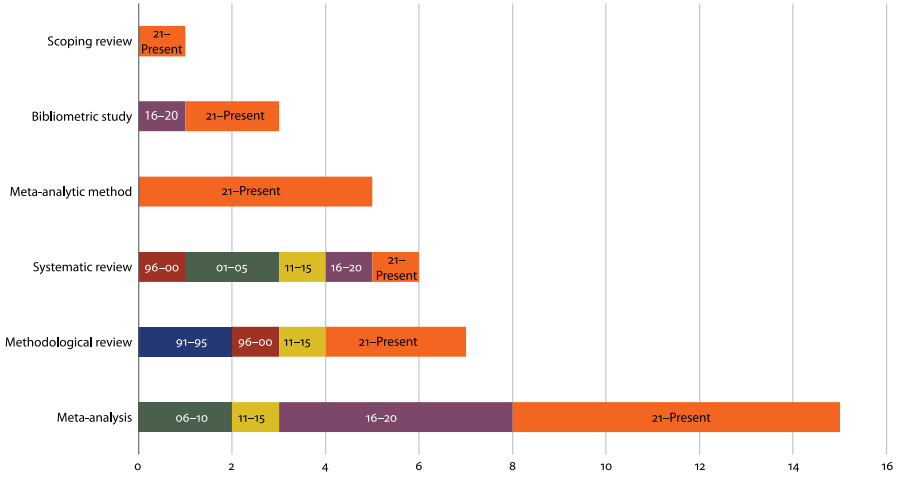


Figure 2. Number of meta-studies by type across five-year periods
Note: The final period includes only through August 2025.

SSLA's research synthesis articles also show a wide coverage of SLA domains (see Figure 3). A focus on L2 skills has been prominent since the 1990s, including reading (e.g., Avery & Marsden, 2019; Hamada et al., 2024), grammar and morphosyntax (e.g., Bardovi-Harlig, 1999; Biber & Reppen, 2002), vocabulary and lexis (e.g., Yanagisawa et al., 2020; Yi & Zhong, 2024), pronunciation and phonology (e.g., Chau & Huensch, 2025; Demir & Kartal, 2022), writing (e.g., Li, 2023), and pragmatics (e.g., Kasper & Dahl, 1991). Research practices represent the second most common theme, with 16 articles, including work on SLA research quality (e.g., Boers et al., 2021; Plonsky, 2013), methodological synthesis in SLA (e.g., Bryfonski et al., 2024; Hulstijn, 1997), and measurement validity (e.g., Li, 2016; Sudina, 2023). Their sustained presence since the 1990s shows the field's emphasis on rigorous research. Instructional variables are another consistent focus, covering topics such as input-based instruction (e.g., Ellis, 2002; Lee & Huang, 2008), strategy training (e.g., Jin & Webb, 2024), and material development (e.g., Biber & Reppen, 2002). Cognitive processes have been synthesized in six studies, including lexical processing (e.g., Rice & Tokowicz, 2020), morphosyntactic processing (e.g., Avery & Marsden, 2019), and attention (e.g., Ellis & Sagarra, 2011). Finally, individual differences (e.g., In'nami et al., 2022; Teimouri et al., 2019) and assessment (e.g., Kostromitina & Plonsky, 2022; Zhao & Aryadoust, 2025) have gained more prominence in the past decade, while learning contexts are receiving less attention recently (e.g., Nunan, 1991; Ziegler, 2016).

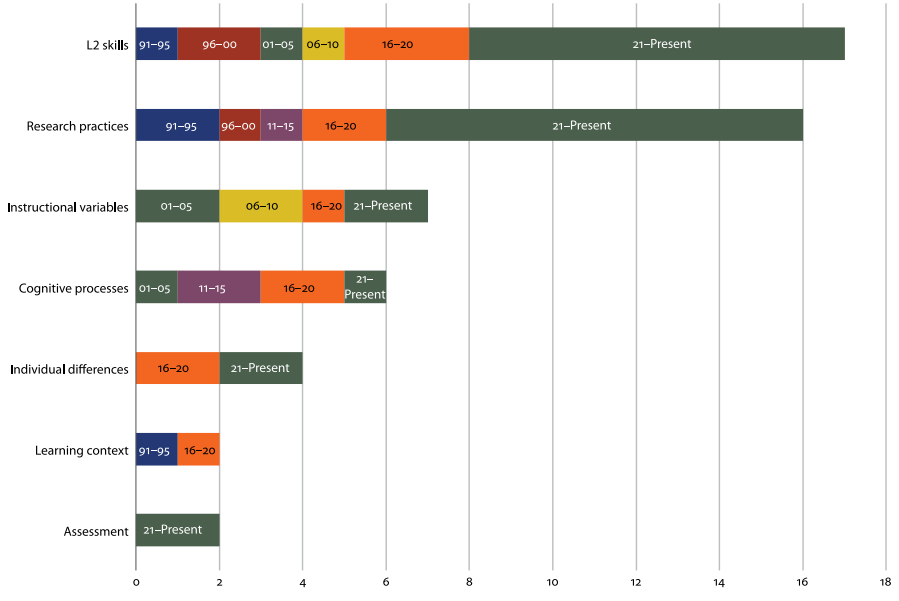


Figure 3. Number of research syntheses by topic across five-year period
Note: The final period includes only through August 2025.

3. Quality in SSLA meta-analyses

In the previous section, we described the frequency of synthetic research found in SSLA both overall and over time. We also presented a summary of the types and substantive foci of synthetic research found in the journal. We now address the quality in a subset of those studies. Specifically, we present the findings from an evaluation of meta-analytic methods found in SSLA.

Method

The instrument used in this mini review was developed by Hu (2024) and consists of 51 items in total. The core sections of the instrument include 39 items organized according to the structure of a meta-analytic report as suggested by Cooper (2016), with each item addressing a specific aspect of meta-analytic methodology or reporting practice. All 39 items were measured on either a three-point Likert scale (*no* = 0, *somewhat* = 1, *yes* = 2) or a binary scale (*no* = 0, *yes* = 2). After coding all items, the scores for every item were summed up to produce a composite score representing the overall quality of each meta-analysis in the sample. The maximum possible score for a meta-analysis is 78. A detailed description of the instrument is provided in Hu (2024). (See Chong, 2025, for an example of another instrument for assessing the quality of research syntheses in Applied Linguistics.)

Description of studies

The first meta-analysis in SSLA was published in 2008. By the end of 2023, a total of 10 meta-analyses had been published and thereby identified for this mini review (see Appendix for a full list). The topics of the reviewed meta-analyses published in SSLA varied, including visual input enhancement, corrective feedback, computer-assisted language learning, individual differences (e.g., aptitude, anxiety, and working memory), self-paced reading, glossing, assessment, and vocabulary learning.

All 10 meta-analyses reported information on the type of effect sizes used and the total number of studies included, whereas only four studies reported the number of effect sizes included. On average, 46.90 studies were included in the 10 meta-analyses (Min = 14; Max = 97). Of the selected meta-analyses, four used *d*-family effect sizes (i.e., standardized mean differences), four used *r*-family effect sizes (i.e., correlations), and two used percentages as an effect size in their meta-analyses.

Overall, the average quality score of the 10 meta-analyses was 41.6 out of a maximum possible 78 (Min = 25; Max = 51; *SD* = 7.41). As shown in Table 1, the

overall quality of the SSLA sub-sample is higher than that of the full sample, except for Step 4. The following sections present detailed results for each step.

Table 1. Mean quality scores for the full sample and the SSLA sub-sample

	Mean (all)	%	Mean (SSLA)	%	Max. possible
Step 1: Formulating the problem ($k=3$)	4.8	80	5.2	87	6
Step 2: Locating literature ($k=5$)	5.6	56	6.7	67	10
Step 3: Gathering information ($k=6$)	5.1	43	6.7	56	12
Step 4: Analyzing the literature ($k=11$)	7.6	35	7.4	34	22
Step 5: Presenting the results ($k=4$)	4.1	51	5.1	64	8
Step 6: Interpreting the evidence ($k=7$)	5.3	38	6.5	46	14
Transparency practices ($k=3$)	3.0	50	4	67	6
Total (k=39)	**35.4**	**45**	**41.6**	**53**	**78**

Note. % = mean score/maximum possible score

Evaluating step 1: Formulating the problem

In step 1, the primary goal is to define the research problem under investigation, including the definitions of variables of interest and the rationale for studying their relationships. The results of this step were overall encouraging, with an average score of 5.2 (maximum possible = 6; $SD = 1.03$).

Looking more closely at each item, all the studies received full scores on the first item that addressed the statement of research questions or relations and the second item on problem justification, indicating that they provided solid justifications for carrying out their meta-analyses and clear statements of the research questions under investigation. Regarding the third item on the rationale for the selection of moderators, four studies failed to do so, and the remaining six studies provided sufficient rationale.

Evaluating step 2: Locating the literature

In step 2, the primary goal is to conduct an exhaustive search of both unpublished (i.e., grey literature) and published literature to be included in the meta-analysis. The average score for this step was 6.7 out of a maximum possible score of 10 (Min = 3; Max = 8; $SD = 1.80$), suggesting that, overall, the quality of this step was less encouraging than step 1 and showed considerable variation across the 10 meta-analyses.

When examining the item on the exhaustiveness of the literature search, only three studies received a code of 2, indicating that the authors carried out a relatively exhaustive search by consulting more than four channels to locate relevant literature. In contrast, the remaining seven studies were coded as 1, as they relied on fewer than four search channels and were therefore considered less comprehensive.

With regard to the inclusion of unpublished literature (e.g., unpublished manuscripts, conference presentations), only three studies reported including two or more types. One study did not include any unpublished literature, whereas the remaining six studies included just one type of report.

All 10 meta-analyses reported relevant inclusion and exclusion criteria. Among them, five were coded as 2, indicating that the criteria were described clearly and comprehensively.

All meta-analyses in the SSLA subsample made their search terms available in the text. However, none of them included Boolean operators.

The final item in this step concerned the reporting of the ending time of the literature search. Of the 10 meta-analyses, eight specified an end date, while the other two did not provide this information.

Evaluating step 3: Gathering the information

The overall goal of step 3 is to code relevant information from the included primary studies identified in the previous step and to evaluate the reliability of the coding results by involving one or more additional coders. The average rating of this step was 6.7 (Min = 4; Max = 8; *SD* = 2.0), out of a maximum possible score of 12.

Notably, all 10 meta-analyses clearly provided definitions and possible values for the coding categories. Regarding coding agreement, six studies reported an overall reliability score, while only one study assessed reliability for each individual item in the coding scheme. The resolution of coding disagreement was also examined, with six meta-analyses providing details on the resolution process. Among the meta-analyses that reported coding agreement values, only four reported information on the training of additional coders.

Evaluating step 4: Analyzing the literature

The main task of step 4 is to select the most appropriate method for aggregating the results of the included primary studies. Overall, the ratings for this step were fairly heterogeneous (maximum possible score = 22; mean = 7.4; Min = 1; Max = 12; *SD* = 3.64).

Half of the meta-analyses explicitly discussed the weighting methods of effect sizes, with two employing a more appropriate approach (e.g., weighting by the reverse of sampling error), while three used a more common but less rigorous approach (i.e., weighting by sample size).

We then turn to effect size dependencies (e.g., multiple effect sizes from a single sample and/or study), addressed by six of the ten meta-analyses in the SSLA sample. Among them, three applied a more appropriate approach (e.g., using a modeling method), while the remaining three employed a less robust strategy (e.g., averaging multiple effect sizes into a single value). In total, three meta-analyses explicitly indicated the choice of models (i.e., fixed-, random-, or mixed-effects models). Among them, two provided justifications of their choices.

Concerning practices related to heterogeneity, three studies examined heterogeneity in observed effect sizes. In addition, nine studies conducted moderator analyses, and eight of them did so appropriately (i.e., without converting continuous moderator variables into categorical ones). In regard to practices related to outliers, four studies reported identifying outliers and handling them appropriately. Unfortunately, none of the meta-analyses reported assessing the study quality of the included primary studies. However, two meta-analyses addressed missing data by contacting the original authors.

Finally, practices related to publication bias were evaluated. Overall, eight meta-analyses reported assessing publication bias. Of these, four employed two or more techniques, while the remaining four used only one technique.

Evaluating step 5: Presenting the results

The primary goal of this step is to thoroughly and faithfully report relevant meta-analytic results. The mean rating was 5.1 out of a maximum possible score of eight (Min = 2; Max = 7; $SD = 1.85$).

First, half of the meta-analyses provided an overall description of the sampled primary studies. Among these, three offered a more thorough description, while the remaining two provided only a brief overview. Most studies employed visualization techniques to present results ($k = 8$), but only six interpreted these visualizations, with just one providing a sufficiently thorough interpretation.

The results for the item on reporting consideration of error were encouraging: All 10 meta-analyses accounted for error around meta-analytic point estimates (e.g., confidence intervals or standard errors).

Evaluating step 6: Interpreting the evidence

The primary function of this stage is to provide contextual, theoretical, and/or practical interpretations of meta-analytic findings. Authors are also expected to interpret findings in relation to their conclusiveness and generalizability. Additionally, both substantive and methodological implications, as well as limitations, should be addressed in this step. Across the sample of meta-analyses published in SSLA, the average rating for this stage was 6.5 out of a maximum possible score of 14 (Min = 4; Max = 11; *SD* = 2.17).

First of all, all meta-analyses provided contextual interpretations of their findings. However, the majority ($k = 6$) limited their interpretation to comparing effect sizes with other related findings or established benchmarks, without taking a step further, for example, to make a practical interpretation of the significance of the effects or relationships.

Among the 10 meta-analyses, four discussed the potential impact of publication bias, only one addressed the potential impact of missing data, and none considered the generalizability of their meta-analytic results.

All of the studies discussed substantive implications, and six provided discussions related to methodological implications. Finally, only four discussed the limitations of their studies.

Evaluating transparency practices

The final section of the coding scheme addressed meta-analytic transparency, covering three specific elements thereof: (1) whether the coding scheme was made available, (2) whether the complete meta-analytic dataset was accessible, and (3) whether a list of the included primary studies was provided. The average score for these practices was 4 out of a maximum possible score of 6 (Min = 3; Max = 6; *SD* = 0.82).

Among the 10 meta-analyses, seven made their coding schemes available either within the paper or online. Two studies provided the complete dataset, and two presented partial datasets. Finally, all the meta-analyses in the SSLA sample provided a list of the primary studies included in their meta-analyses.

4. Summary and conclusion

This review — a kind of synthesis itself (specifically, a second-order review/synthesis, or a meta-review) — has shown SSLA to be a productive venue for the dissemination of synthetic research in Applied Linguistics. SSLA has published

syntheses across a range of topics and traditions within L2 research. Importantly, it has also maintained a level of meta-analytic rigor that surpasses what is typical in the field.

It is also worth noting that many of the field's advances in synthetic and meta-analytic methods can be found in SSLA. Some of these innovative techniques are found in individual articles that focus explicitly on the meta-analytic methods themselves. For example, Norouzian (2021) illustrated the need for and provided guidance on more appropriate estimates of interrater reliability in the case of categorical variables in meta-analytic coding schemes (i.e., the *S* index). Lindstromberg (2022) described one particular technique for assessing publication bias in a given body of primary studies. And Norouzian and Bui (2024) addressed techniques for modelling the multivariate and multi-level datasets often found in L2 meta-analyses. The impact of these articles is already evident. As two cases in point, Sudina (2023) calculated and reported the *S* index described in Norouzian (2021), and the modelling techniques introduced in Norouzian and Bui (2024) were taken up very shortly thereafter by Huang and Zhang (2025). Still other instances of innovation in meta-analytic methods can be found elsewhere in SSLA's body of meta-analytic work. Several authors have meta-analysed reliability estimates in what is referred to as reliability generalization meta-analysis (e.g., In'nami et al., 2022; Sudina, 2023; Zhao & Aryadoust, 2025). Others have likewise sought to address reliability in the context of otherwise substantively-focused syntheses by correcting meta-analytic outcomes for attenuation due to measurement error in primary studies (e.g., Kostromitina & Plonsky, 2022; Teimouri et al., 2019). And at least one meta-analysis in SSLA has sought to reach a wider audience by reporting meta-analytic outcomes not only in standardized mean difference effect sizes but using what are often referred to as 'common sense' effect sizes (Ziegler, 2016).

Despite these innovations and advances, neither SSLA nor any other journal should be complacent. There is no such thing as a perfect study, and the body of secondary research in SSLA could certainly be improved, expanded, replicated, and re-analyzed.

Looking forward, it is hard to know what role SSLA will play in the future of synthetic L2. We hope, however, to see continued activity and effort in a few different directions. First, as we just alluded to, there is ample room for meta-analytic replication studies, which serve to expand and update our knowledge in a given domain. Such studies can also improve on the initial study's methods — perhaps employing one or more novel developments (e.g., corrections for attenuation due to measurement error) to arrive at more trustworthy, stable, and actionable outcomes (see Zhao & Plonsky, 2024). We also hope and expect to see additional applications of bibliometric methods, which have begun to appear in SSLA and

elsewhere (Plonsky, 2023; Zhang & Aryadoust, 2025), but which have yet to be employed in most domains across the field. Finally, as we exemplified here, and as found in several recent works (e.g., Chalmers et al., 2024; Hu, 2024; Vuogan & Li, 2024), we also hope and expect to see secondary research in SSLA that embraces the research synthesis and, more broadly meta-science agenda to quality and its many facets at the primary and secondary levels (Plonsky, 2024). One way to move in this direction, and one related goal for SSLA, will be to develop journal- and field-specific guidelines for the conduct of secondary research in Applied Linguistics.

AI use declaration statement

The authors confirm that no AI tools were used in the preparation, writing, formatting, and editing of this work.

References

Avery, N., & Marsden, E. (2019). A meta-analysis of sensitivity to grammatical information during self-paced reading: Towards a framework of reference for reading time effect sizes. *Studies in Second Language Acquisition*, *41*(5), 1055–1087.

Bardovi-Harlig, K. (1999). From morpheme studies to temporal semantics: Tense-aspect research in SLA. *Studies in Second Language Acquisition*, *21*(3), 341–382.

Biber, D., & Reppen, R. (2002). What does frequency have to do with grammar teaching? *Studies in Second Language Acquisition*, *24*(2), 199–208.

Boers, F., Bryfonski, L., Faez, F., & McKay, T. (2021). A call for cautious interpretation of meta-analytic reviews. *Studies in Second Language Acquisition*, *43*, 2–24.

Bryfonski, L., Ku, Y., & Mackey, A. (2024). Research methods for IDs and TBLT: A substantive and methodological review. *Studies in Second Language Acquisition*, *46*(3), 617–643.

Chalmers, H., Brown, J., & Koryakina, A. (2024). Topics, publication patterns, and reporting quality in systematic reviews in language education: Lessons from the international database of education systematic reviews (IDESR). *Applied Linguistics Review*, *15*, 1645–1669.

Chau, T., & Huensch, A. (2025). The relationships among L2 fluency, intelligibility, comprehensibility, and accentedness: A meta-analysis. *Studies in Second Language Acquisition*, *47*(1), 282–307.

Chong, S.W. (2025). Synthesis Methods and Reporting Tool (SMART) for research syntheses in Applied Linguistics. *Research Synthesis in Applied Linguistics*, *1*(1), 17–38.

Chong, S.W., & Plonsky, L. (2024). A typology of secondary research in Applied Linguistics. *Applied Linguistics Review*, *15*(4), 1569–1594.

Cooper, H. (2016). *Research synthesis and meta-analysis: A step-by-step approach*. Sage.

Demir, Y., & Kartal, G. (2022). Mapping research on L2 pronunciation: A bibliometric analysis. *Studies in Second Language Acquisition*, *44*(5), 1211–1239.

Ellis, N. C., & Sagarra, N. (2011). Learned attention in adult language acquisition: A replication and generalization study and meta-analysis. *Studies in Second Language Acquisition*, *33*(4), 589–624.

Ellis, R. (2002). Does form-focused instruction affect the acquisition of implicit knowledge? A review of the research. *Studies in Second Language Acquisition*, *24*(2), 223–236.

Hamada, A., Shimizu, H., Hoshino, Y., Takaki, S., & Ushiro, Y. (2024). Robust evidence for the simple view of second language reading: Secondary meta-analysis of Jeon and Yamashita (2022). *Studies in Second Language Acquisition*, *46*(5), 1355–1372.

Hu, Y. (2024). Evaluating methodological and reporting practices of L2 meta-analyses: A second-order review (Unpublished doctoral dissertation). Northern Arizona University.

Huang, F., & Zhang, H. (2025). The relationship between boredom and second language achievement: A multilevel meta-analysis. *Studies in Second Language Acquisition*, *47*, 488–511.

Hulstijn, J. H. (1997). Second language acquisition research in the laboratory: Possibilities and limitations. *Studies in Second Language Acquisition*, *19*(2), 131–143.

In'nami, Y., Hijikata, Y., & Koizumi, R. (2022). Working memory capacity and L2 reading: A meta-analysis. *Studies in Second Language Acquisition*, *44*(2), 381–406.

Jin, Z., & Webb, S. (2024). The effectiveness of note taking through exposure to L2 input: A meta-analysis. *Studies in Second Language Acquisition*, *46*(2), 404–426.

Kasper, G., & Dahl, M. (1991). Research methods in interlanguage pragmatics. *Studies in Second Language Acquisition*, *13*(2), 215–247.

Kostromitina, M., & Plonsky, L. (2022). Elicited imitation tasks as a measure of L2 proficiency: A meta-analysis. *Studies in Second Language Acquisition*, *44*(3), 886–911.

Lee, S.-K., & Huang, H.-T. (2008). Visual input enhancement and grammar learning: A meta-analytic review. *Studies in Second Language Acquisition*, *30*(3), 307–331.

Li, S. (2016). The construct validity of language aptitude: A meta-analysis. *Studies in Second Language Acquisition*, *38*(4), 801–842.

Li, S. (2023). Working memory and second language writing: A systematic review. *Studies in Second Language Acquisition*, *45*(3), 647–679.

Lindstromberg, S. (2022). P-curving as a safeguard against p-hacking in SLA research: A case study. *Studies in Second Language Acquisition*, *44*, 1155–1180.

Norouzian, R. (2021). Interrater reliability in second language meta-analyses: The case of categorical moderators. *Studies in Second Language Acquisition*, *43*, 896 — 915.

Norouzian, R., & Bui, G. (2024). Meta-analysis of second language research with complex research designs. *Studies in Second Language Acquisition*, *46*, 251 — 276.

Norris, J. M., & Ortega, L. (2000). Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis. *Language Learning*, *50*, 417–528.

Nunan, D. (1991). Methods in second language classroom-oriented research: A critical review. *Studies in Second Language Acquisition*, *13*(2), 249–274.

Plonsky, L. (2012). Replication, meta-analysis, and generalizability. In G. Porte (Ed.), *Replication research in Applied Linguistics* (pp. 116–132). Cambridge University Press.

Plonsky, L. (2013). Study quality in SLA: An assessment of designs, analyses, and reporting practices in quantitative L2 research. *Studies in Second Language Acquisition*, *35*(4), 655–687.

Plonsky, L. (2023). *Bibliometrics in Applied Linguistics*. Special issue in *Studies in Second Language Learning and Teaching*, *13*.

Plonsky, L. (2024). Study quality as an intellectual and ethical imperative: A proposed framework. *Annual Review of Applied Linguistics*, *44*, 4–18.

Plonsky, L., Hu, Y., Sudina, E., & Oswald, F.L. (2023). Advancing meta-analytic methods in L2 research. In A. Mackey & S. Gass (Eds.), *Current approaches in second language acquisition research* (pp. 304–333). Wiley Blackwell.

Rice, C.A., & Tokowicz, N. (2020). A review of laboratory studies of adult second language vocabulary training. *Studies in Second Language Acquisition*, *42*(2), 439–470.

Sudina, E. (2023). Scale quality in second-language anxiety and WTC: A methodological synthesis. *Studies in Second Language Acquisition*, *45*(5), 1427–1455.

Teimouri, Y., Goetze, J., & Plonsky, L. (2019). Second language anxiety and achievement: A meta-analysis. *Studies in Second Language Acquisition*, *41*(2), 363–387.

Vuogan, A., & Li, S. (2024). A systematic review of meta-analyses in second language research: Current practices, issues, and recommendations. *Applied Linguistics Review*, *15*(4), 1621–1644.

Yanagisawa, A., Webb, S., & Uchihara, T. (2020). How do different forms of glossing contribute to L2 vocabulary learning from reading? A meta-regression analysis. *Studies in Second Language Acquisition*, *42*(2), 411–438.

Yi, W., & Zhong, Y. (2024). The processing advantage of multiword sequences: A meta-analysis. *Studies in Second Language Acquisition*, *46*(2), 427–452.

Zhao, H., & Aryadoust, V. (2025). A meta-analysis of the reliability of second language reading comprehension assessment tools. *Studies in Second Language Acquisition*, *47*(1), 388–416.

Zhao, S., & Plonsky, L. (2024). Reliability: What do we mean when we (don't) talk about it? *Estudios de Lingüística Inglesa Aplicada (ELIA)*, *24*, 245–258.

Ziegler, N. (2016). Synchronous computer-mediated communication and interaction: A meta-analysis. *Studies in Second Language Acquisition*, *38*(3), 553–586.

Appendix

List of research synthesis articles published in SSLA. Studies included in the quality assessment were labelled with an asterisk.

*Avery, N., & Marsden, E. (2019). A meta-analysis of sensitivity to grammatical information during self-paced reading: Towards a framework of reference for reading time effect sizes. *Studies in Second Language Acquisition*, *41*(5), 1055–1087.

Boers, F., Bryfonski, L., Faez, F., & McKay, T. (2021). A call for cautious interpretation of meta-analytic reviews. *Studies in Second Language Acquisition*, *43*(1), 2–24.

Bardovi-Harlig, K. (1999). From morpheme studies to temporal semantics: Tense-aspect research in SLA. *Studies in Second Language Acquisition*, *21*(3), 341–382.

Biber, D., & Reppen, R. (2002). What does frequency have to do with grammar teaching? *Studies in Second Language Acquisition*, *24*(2), 199–208.

Bryfonski, L., Ku, Y., & Mackey, A. (2024). Research methods for IDs and TBLT: A substantive and methodological review. *Studies in Second Language Acquisition*, *46*(3), 617–643.

Chau, T., & Huensch, A. (2025). The relationships among L2 fluency, intelligibility, comprehensibility, and accentedness: A meta-analysis. *Studies in Second Language Acquisition*, *47*(1), 282–307.

Demir, Y., & Kartal, G. (2022). Mapping research on L2 pronunciation: A bibliometric analysis. *Studies in Second Language Acquisition*, *44*(5), 1211–1239.

Ellis, N. C., & Sagarra, N. (2011). Learned attention in adult language acquisition: A replication and generalization study and meta-analysis. *Studies in Second Language Acquisition*, *33*(4), 589–624.

Ellis, R. (2002). Does form-focused instruction affect the acquisition of implicit knowledge? : A review of the research. *Studies in Second Language Acquisition*, *24*(2), 223–236.

Hamada, A., Shimizu, H., Hoshino, Y., Takaki, S., & Ushiro, Y. (2024). Robust evidence for the simple view of second language reading: Secondary meta-analysis of Jeon and Yamashita (2022). *Studies in Second Language Acquisition*, *46*(5), 1355–1372.

Hiver, P., Al-Hoorie, A. H., & Evans, R. (2022). Complex dynamic systems theory in language learning: A scoping review of 25 years of research. *Studies in Second Language Acquisition*, *44*(4), 913–941.

Hulstijn, J. H. (1997). Second language acquisition research in the laboratory: Possibilities and limitations. *Studies in Second Language Acquisition*, *19*(2), 131–143.

*In'nami, Y., Hijikata, Y., & Koizumi, R. (2022). Working memory capacity and L2 reading: A meta-analysis. *Studies in Second Language Acquisition*, *44*(2), 381–406.

Jin, Z., & Webb, S. (2024). The effectiveness of note taking through exposure to L2 input: A meta-analysis. *Studies in Second Language Acquisition*, *46*(2), 404–426.

Kasper, G., & Dahl, M. (1991). Research methods in interlanguage pragmatics. *Studies in Second Language Acquisition*, *13*(2), 215–247.

*Kostromitina, M., & Plonsky, L. (2022). Elicited imitation tasks as a measure of L2 proficiency: A meta-analysis. *Studies in Second Language Acquisition*, *44*(3), 886–911.

*Lee, S.-K., & Huang, H.-T. (2008). Visual input enhancement and grammar learning: A meta-analytic review. *Studies in Second Language Acquisition*, *30*(3), 307–331.

*Li, S. (2016). The construct validity of language aptitude: A meta-analysis. *Studies in Second Language Acquisition*, *38*(4), 801–842.

Li, S. (2023). Working memory and second language writing: A systematic review. *Studies in Second Language Acquisition*, *45*(3), 647–679.

Lindstromberg, S. (2022). P-curving as a safeguard against p-hacking in SLA research: A case study. *Studies in Second Language Acquisition*, *44*, 1155–1180.

*Lyster, R., & Saito, K. (2010). Oral feedback in classroom SLA: A meta-analysis. *Studies in Second Language Acquisition*, *32*(2), 265–302.

Norouzian, R. (2021). Interrater reliability in second language meta-analyses: The case of categorical moderators. *Studies in Second Language Acquisition*, *43*, 896 — 915.

Norouzian, R., & Bui, G. (2024). Meta-analysis of second language research with complex research designs. *Studies in Second Language Acquisition*, *46*, 251 — 276.

Nunan, D. (1991). Methods in second language classroom-oriented research: A critical review. *Studies in Second Language Acquisition*, *13*(2), 249–274.

Plonsky, L. (2013). Study quality in SLA: An assessment of designs, analyses, and reporting practices in quantitative L2 research. *Studies in Second Language Acquisition*, *35*(4), 655–687.

Rice, C.A., & Tokowicz, N. (2020). A review of laboratory studies of adult second language vocabulary training. *Studies in Second Language Acquisition*, *42*(2), 439–470.

Sudina, E. (2023). Scale quality in second-language anxiety and WTC: A methodological synthesis. *Studies in Second Language Acquisition*, *45*(5), 1427–1455.

*Teimouri, Y., Goetze, J., & Plonsky, L. (2019). Second language anxiety and achievement: A meta-analysis. *Studies in Second Language Acquisition*, *41*(2), 363–387.

Tolentino, L.C., & Tokowicz, N. (2011). Across languages, space, and time: A review of the role of cross-language similarity in L2 (morpho)syntactic processing as revealed by fMRI and ERP methods. *Studies in Second Language Acquisition*, *33*(1), 91–125.

Vitta, J., Nicklin, C., & McLean, S. (2022). Effect size-driven sample-size planning, randomization, and multisite use in L2 instructed vocabulary acquisition experimental samples. *Studies in Second Language Acquisition*, *44*, 1424–1448.

*Yanagisawa, A., Webb, S., & Uchihara, T. (2020). How do different forms of glossing contribute to L2 vocabulary learning from reading? A meta-regression analysis. *Studies in Second Language Acquisition*, *42*(2), 411–438.

*Yanagisawa, A., & Webb, S. (2022). Involvement Load Hypothesis Plus: Creating an improved predictive model of incidental vocabulary learning. *Studies in Second Language Acquisition*, *44*(5), 1279–1308.

Yi, W., & Zhong, Y. (2024). The processing advantage of multiword sequences: A meta-analysis. *Studies in Second Language Acquisition*, *46*(2), 427–452.

Zhang, S., & Aryadoust, V. (2025). Why are some articles highly cited in Applied Linguistics? A bibliometric study. *Studies in Second Language Acquisition*, *47*(1), 3–25.

Zhang, X. (2020). A bibliometric analysis of second language acquisition between 1997 and 2018. *Studies in Second Language Acquisition*, *42*(1), 199–222.

Zhao, H., & Aryadoust, V. (2025). A meta-analysis of the reliability of second language reading comprehension assessment tools. *Studies in Second Language Acquisition*, *47*(1), 388–416.

*Ziegler, N. (2016). Synchronous computer-mediated communication and interaction: A meta-analysis. *Studies in Second Language Acquisition*, *38*(3), 553–586.

CHAPTER 4

Planning, doing, and disseminating research syntheses

An introduction to the new journal, *Research Synthesis in Applied Linguistics*

Sathena Chan & Sin Wang Chong
University of Bedfordshire | University of St Andrews

This chapter introduces a newly launched international journal on research synthesis in Applied Linguistics, published by Taylor & Francis. As the editors-in-chief, we reflect on the process of establishing the first journal in Applied Linguistics dedicated to publishing research syntheses. We outline the journal's aims, scope, and submission categories, and share our vision for how it would contribute to advancing discussions and debates on research synthesis methodologies within and beyond the field. In particular, we detail initiatives to promote methodological rigour of research syntheses through pre-registration, use of a field-specific methodological guide, and open-access infographics dissemination.

Keywords: research synthesis, applied linguistics, SMART, pre-registration, infographics

From recognising gaps to founding RSAL

Research syntheses, or evidence syntheses, are a type of secondary research that employs systematic, transparent, reproducible, and often pre-registered methodologies to synthesise published research (Chong & Plonsky, 2024). The focus of research syntheses can be substantive (e.g., language testing, translanguaging) and/or methodological (e.g., narrative inquiry).

In the field of Applied Linguistics, meta-analyses are traditionally the predominant form of research synthesis (Chong, 2025). Researchers have found it challenging to publish other emergent types of research syntheses, especially those that employ qualitative methodologies (e.g., qualitative research synthesis, scoping review, systematic literature review). The challenges of publishing

https://doi.org/10.1075/rmal.17.04cha
Available under the CC BY-NC-ND 4.0 license.© 2026 John Benjamins Publishing Company

research syntheses in reputable journals are partly due to an inconsistent use of terminology and the lack of a common understanding of the various types of research syntheses in Applied Linguistics. Secondly, the typical word limit between 8000–10000 words of most journals in the field unintentionally makes it harder for qualitative research synthesis than meta-analyses to be published. Thirdly, the discussion about field-specific standards for conducting and reporting research syntheses has been scarce. Questions such as the minimum quality requirement of research syntheses and how research synthesis results should be interpreted in Applied Linguistics remain to be explored.

Through our previous executive committee roles at the British Association for Applied Linguistics (BAAL), we exchanged some of these challenges faced by BAAL special interest group members who were interested in conducting and/or publishing synthetic research. We shared an aspiration of starting a journal specifically for research syntheses to provide a dedicated platform for the publication of rigorously conducted research syntheses of all types in Applied Linguistics. However, we were not sure if such needs for a new methodological journal were shared and recognised by other researchers and editors in the field. So, we asked. We were encouraged by the positive feedback and constructive suggestions generously offered by our colleagues. Sin Wang convened a working group to iron out the ideas. We then put together a proposal highlighting the increasing popularity of research syntheses in Applied Linguistics and the aforementioned challenges of publishing synthetic research. We explained why such a methodological journal would be beneficial for the development of Applied Linguistics and our goals for the new journal. In the process of preparing the journal proposal, our aspiration became tangible. After several rounds of review and feedback, we were grateful that our proposal was approved by Taylor & Francis, one of the leading publishers. We also put together an editorial board. With the support of the publisher and our editorial board, the details of the proposal were thoroughly discussed and refined. The new journal Research Synthesis in Applied Linguistics (RSAL) was founded in 2024 and officially launched on 13 January 2025.

About RSAL

The journal webpage (https://www.tandfonline.com/journals/rsal20) sets out the scope and aims of the RSAL:

> RSAL is an international journal publishing high-quality, original research. This journal is methodological in focus, so it welcomes research syntheses on all topics in Applied Linguistics, but especially those that have a strong relevance to

> professional practice (e.g., language teaching, learning, and assessment). Adhering to the definition of research synthesis in the "Aims and Scope" page, this journal does not publish "traditional literature reviews" that do not include a pre-determined methodology, i.e., search strategy, inclusion and/or exclusion criteria, screening of documents, data extraction, data synthesis. Examples of "traditional literature reviews" include narrative reviews, critical reviews etc. (Chong & Plonsky, 2023). Following good practices of conducting research synthesis, preregistration is strongly encouraged for all syntheses published in this journal, i.e., accompanied by a protocol published either with the journal or on a research repository (e.g., IDESR, OSF). The journal particularly welcomes submissions that employ innovative and new methods/approaches to synthesising evidence and research (e.g., bibliometric review, second-order review).

Publication types

It is our mission to welcome publication of high-quality research syntheses on any topic (methodological and substantive) in Applied Linguistics. RSAL accepts four types of articles: *Research Synthesis, Methodological Article, Research Synthesis Protocol* and *Book Review*.

For *Research Synthesis* papers, we offer a generous word count up to 20,000 words, a feature welcomed by many researchers who aim to publish qualitative and large-scale research syntheses. The extended length allows for the depth, nuance and complexity required to present extensive and in-depth findings with evidence, e.g., participant quotes, narrative of specific contexts, and examples of thematic analysis.

To promote consistency of research syntheses in the field, we require authors to indicate clearly the type of research synthesis they belong to in their title, abstract, and methodology by referring to published typologies such as Chong and Plonsky (2023).

Methodological Articles are papers that focus on a methodological issue or method related to conducting research syntheses. We allow either a descriptive or discursive focus for these papers. Descriptive papers, for example, can introduce a new method or tool for conducting a type of research synthesis. Discursive papers, on the other hand, tend to discuss a particular methodological issue, e.g., use of generative AI tools in conducting research syntheses. We ask that authors include examples, preferably in Applied Linguistics or a cognate field, to illustrate the information and arguments presented. Methodological papers can be commissioned by the editors-in-chief and might take the form of shorter communication pieces.

Research Synthesis Protocols are study plans of research syntheses. Research synthesis protocols will be reviewed by one of the journal editors. We will discuss the importance of protocols and pre-registration more broadly in the next section.

Book Reviews are reviews of books relevant to research syntheses in Applied Linguistics. Our Book Review Editor commissions book reviews and looks after this article type. Book reviews are also peer-reviewed.

Together, the four publication types cover the range of publications relevant to research syntheses and provide a good balance of disseminating synthetic research and shaping theoretical/methodological discussions in our field. We are scheduled to publish two issues each year, with 10 publications in each issue. Our inaugural issue with ten publications was published in August 2025.

Editorial Board

Currently, the editorial board of RSAL includes a book review editor, bibliography editors, editorial advisory board, technical advisory board, editorial board and a graphical abstract designer.

Our aim is to include international members with different levels of experience in research syntheses as well as the varied substantive expertise on our editorial board. Bibliography editors, with our support, are responsible for collating a bibliography of all research synthesis publications in Applied Linguistics periodically and reporting the general trends and developments of research syntheses in our field. It is hoped that over time this section would provide a useful reference resource for students and colleagues interested in research syntheses. We give priorities to earlier-career researchers (e.g., doctoral research students and those who have had their PhD within five years) when appointing Bibliography Editors. Graphical abstract designer who looks after the design of our graphical abstracts is another role we give priorities to earlier-career researchers.

The editorial advisory board are experienced editors and/or reviewers of other journals and leading researchers in our field. We have benefited from their feedback on drafts of RSAL documentation and constructive advice on different aspects of operating a new journal. As a methodological journal, we also include a technical advisory board of colleagues who are experts in various research synthesis methods in other fields. The editorial board are colleagues who contribute their time regularly to review manuscripts. Our aim is for all research syntheses to be reviewed by two reviewers — one reviewer with substantive expertise and the other one with expertise in the type of research synthesis.

Unique features to promote methodological rigour

As the first and only methodological journal focusing on research syntheses in Applied Linguistics, we recognise the multifaceted and pioneering role of RSAL (see Plonsky, Isaacs, & Taylor, 2025, for a thoughtful discussion piece about the opportunities and challenges of the new journal). Early methodological journals often help define what constitutes rigorous research in that field. Among others, RSAL should serve to promote an establishment of methodological standards applicable and relevant to Applied Linguistics. To achieve this, we introduce three initiatives of RSAL that aim to support researchers to publish rigorous research syntheses and disseminate their findings to a wider audience beyond the readers of the journal.

Pre-registration

RSAL offers two routes to pre-registration: title registration and protocol registration. Title registration requires researchers to complete a Google Form responding to four questions: tentative title of the research synthesis, research questions, personnel (team members and their roles), and timeline. This is a low-effort exercise that allows researchers to 'timestamp' their research synthesis ideas. The research syntheses that are title registered are viewable via an online spreadsheet that is linked to the journal website. In this way, researchers in the field can review ongoing research syntheses before starting a new review project. This has several benefits. First, it avoids research waste (Isaacs & Chalmers, 2025) by ensuring that new research syntheses do not duplicate in focus. Second, it encourages collaboration among researchers who have a shared research interest. Instead of working on different research syntheses on similar topics, they can join forces to conduct a more thorough and substantial review on an area of interest. Currently, 40 ongoing research syntheses are title registered with RSAL. Title registration is a well-established practice in the disciplines of Medicine and Healthcare. However, this is less common in our field. In Education, the journal, Campbell Systematic Reviews, supports title registration for research syntheses on a social sciences topic including education. In Applied Linguistics, RSAL is the only platform that offers title registration.

Another route to pre-register with RSAL is through protocol registration. A protocol is a detailed plan of how a research synthesis will be conducted. In effect, it is a draft of the methodology section of a research synthesis manuscript. A protocol submitted to RSAL needs to include the following elements: background and review rationale; review questions; inclusion and exclusion criteria; search

strategy; data extraction and management; study appraisal (if applicable); data synthesis; reporting; personnel; conflict of interest; timeline; references; supplementary materials (if applicable). These components are informed by PRISMA-P (Moher et al., 2015). Completing protocol registration enhances the transparency and reproducibility of a research synthesis. If you submit a protocol to RSAL, it is reviewed by one of the editors-in-chief to provide constructive comments for improving the research design of the research synthesis. Publishing a protocol with RSAL has added benefits. Compared to publishing a protocol on a research repository like the Open Science Framework, accepted protocols submitted to RSAL are published as a journal article with a designated DOI, meaning it is fully citable. Moreover, since RSAL is one of the Taylor & Francis' journals, it is indexed in research databases and search engines, making the protocols more discoverable. By publishing a protocol, researchers get to focus more on reporting findings and insights rather than describing methodological steps in the final article; readers can be directed to read the protocol for details related to the methodology, while the methodology section of the full article can be about key stages and any modification from the original plan.

Synthesis Methods and Reporting Tool (SMART)

RSAL is on a mission to promote critical discussions and innovative thinking about research synthesis methodologies. One obstacle faced by applied linguists when engaging in such conversations is the lack of field-specific guidelines and standards. While guidelines such as PRISMA and AMSTAR2 exist and are widely used in different disciplines, there is a need to develop a methodological guide for conducting research syntheses in Applied Linguistics, given the interdisciplinary and constantly evolving nature of the field. Together with the launch of RSAL, the Synthesis Methods and Reporting Tool (SMART) was introduced (Chong, 2025). SMART is a 22-item checklist that covers three key stages of research synthesis, namely preparation, method, and reporting. Its development is underpinned by four methodological qualities known as STAR: systematicity, transparency, accessibility, and reflexivity. The process of developing SMART is collaborative, involving input from more than 20 researchers in Applied Linguistics at different career stages, representing a wide range of sub-fields, and research synthesis specialists in other disciplines. Unlike other research synthesis guides, SMART addresses the tension between standardisation and innovation. Like all methodological guides, SMART aims to establish methodological thresholds and expectations that work for Applied Linguistics; at the same time, we are reminded by Plonsky et al. (2025) to remain critical, reflecting on key questions

such as: What are these methodological standards? Why are these standards suitable? Who gets to set these standards? What philosophical and epistemological views are advocated through these standards? It is important to understand that Applied Linguistics is an eclectic field of inquiry, straddling disciplinary boundaries of Education, Psychology, Healthcare, Linguistics, and many more. To accommodate the dynamic development of the field, it is essential that tools like SMART would encourage diverse and innovative methodological practices for conducting research syntheses.

How did SMART strike a balance between methodological standardisation and innovation? We do so by introducing baseline practices and preferred practices, which is a unique feature of SMART. Baseline practices refer to actions that are deemed essential to conducting a high-quality research synthesis; for example, the involvement of at least two reviewers in screening, data extraction, and data synthesis is regarded as non-negotiable. Preferred practices, on the other hand, are recommendations recognising methodological practices in other disciplines that improve the rigour of research syntheses. An example is protocol registration, which was discussed in the previous section. The inclusion of baseline and preferred practices encouraged Applied Linguists to consider alternative or additional methodological actions that would enhance the quality of synthetic research. This feature allows SMART to remain agile and responsive to the latest methodological developments in the field. For instance, it is expected that in future versions of SMART, some preferred practices will become baseline practices due to field consensus, while new recommended actions emerge such as how artificial intelligence can be used to conduct research syntheses.

The SMART checklist is currently used by RSAL as a guide to assist prospective authors in planning, implementing, and reporting their research syntheses. It is also used by the editors and peer reviewers to assess submissions and provide feedback. Despite its current use by RSAL, SMART was developed as an educational tool for researchers who are new to research syntheses, who use it to guide their research process; for experienced synthesists, SMART can be referred to when exploring methodological innovations. To make SMART useful to the field, a validation study is being conducted to gather insights into users' experience and the validity of the checklist items. The findings of this study will help refine SMART, making it a more fit-for-purpose tool for those seeking research synthesis support.

Infographics

Research syntheses contain rich insights for researchers, practitioners, and policymakers. Due to the vast volume of information, research syntheses are usually

presented as long reports. In the case of RSAL, research syntheses are published as journal articles that can be up to 20,000 words. While informative, they remain inaccessible to most users of evidence such as teachers. One of the principles that underpins the SMART checklist is *accessibility*, which prompts researchers to consider how synthesised findings can be communicated in a digestible manner to a non-academic audience. There are different ways to make research syntheses more accessible to readers outside of academia, such as using diagrams, figures, and tables. Other examples include summarising findings in plain language, such as evidence briefs on platforms such as OASIS (oasis-database.org).

RSAL is committed to open science practices, especially when it comes to facilitating effective science communication to maximise research impact. RSAL collaborates with TESOLgraphics (tesolgraphics.com), an international platform jointly led by academics and practitioners that produces one-page infographics for research syntheses. These visual summaries are produced by language teachers and educators, ensuring the summarised information remains useful and usable to other professionals in the industry (Sato et al., 2024). When a research synthesis is accepted for publication by RSAL, the authors are asked to provide a textual summary in bullet points following a template. This summary is then passed onto a member of the TESOLgraphics team, who creates an infographic based on the provided content. At the time of writing, RSAL has appointed our own graphic abstract designer. RSAL's infographics usually contain the following sections: definition of key term(s), number of studies summarised, what teachers can learn from the research synthesis, contextual information about the synthesis (e.g., EFL), what the major findings are, and how teachers can turn these findings into action.

The infographics are available to view as a thumbnail on the article webpage as well as on the first page of the publication, before the article title and abstract. In our negotiations with Taylor & Francis, we emphasised the importance of making these infographics accessible to practitioners. We are pleased that we can make the thumbnail of the infographic open-access on RSAL's website, as well as upload a copy of the infographic to the TESOLgraphics site that readers can download without paying a fee. The TESOLgraphics and RSAL teams also share the infographics on social media, reaching language teachers around the world. At the time of writing, the TESOLgraphics X account has a follower base of over 4K, and their website is visited by educators from 126 countries.

Multilingual title, abstract and keywords

To recognise and promote multilingualism in our field, another initiative of RSAL includes the optional publication of translated title, abstract and keywords, for all

article types, except book reviews. Kato and Sato's protocol (2025) is our first publication to include both English and Japanese titles, abstracts and keywords.

Concluding thoughts

RSAL is on a mission to advocate for responsible and ethical production, discussion, and use of research syntheses in the evidence ecosystem to inform evidence generation, facilitate evidence translation, and encourage use of evidence (Shepherd, 2014). As a new methodological journal, we have introduced the various initiatives, including a dual substantive and methodological review system, two routes for pre-registration (title registration and protocols), introduction of SMART as a methodological assessment guide, use of multimodal infographics to promote open-access and reach for non-academic audience, and the recognition of multilingualism through the publication of translated titles, abstracts and keywords. These initiatives will be reviewed periodically as the journal develops.

AI use declaration statement

The authors confirm that no AI tools were used in the preparation, writing, formatting, and editing of this work.

References

Chong, S.W. (2025). Synthesis Methods and Reporting Tool (SMART) for research syntheses in Applied Linguistics. *Research Synthesis in Applied Linguistics*, *1*(1), 17–38.

Isaacs, T., & Chalmers, H. (2025). Reducing 'avoidable research waste' in Applied Linguistics research: Insights from healthcare research. *Language Teaching*, *58*(2), 214–231.

Kato, S., & Sato, M. (2025). Protocol for self-regulated learning in English vocabulary: A scoping review of frameworks and practice. *Research Synthesis in Applied Linguistics*, 1–13.

Moher, D., Shamseer, L., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., Shekelle, P., & Stewart, L.A., & PRISMA-P Group. (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Systematic Reviews*, *4*, 1.

Chong, S. & Plonsky, L. (2024). A typology of secondary research in Applied Linguistics. *Applied Linguistics Review*, *15*(4), 1569–1594.

Plonsky, L., Isaacs, T., & Taylor, L. (2025). Opportunities and challenges for research synthesis in applied linguistics: Three scholars reflect. *Research Synthesis in Applied Linguistics*, *1*(1), 5–16.

Sato, M., Chong, S.W., Aktar, T., Cowell, J., Kong, M.S., & Shaahdadi, M. (2024). Creating and sustaining a platform for researchers and teachers to communicate: An example of *TESOLgraphics*. *Innovation in Language Learning and Teaching*, 1–11.

Shepherd, J. (2014). *How to achieve more effective services: The evidence ecosystem*. What Works Network & Cardiff University. https://orca.cardiff.ac.uk/id/eprint/69077/1/2014_JPS_What_Works.pdf

RSAL title registration form

https://docs.google.com/forms/d/e/1FAIpQLSc08Z07mRrH5VNhP3NHHtoDXZbimUdt-Q1iIjF7HyPHhyVSAw/viewform (to fill in)

https://docs.google.com/spreadsheets/d/1VfMjcHiT-mNPGhos7ts9phXE1sBkJ4yA5bCQ5vxvflc/edit?gid=603690919#gid=603690919 (to view)

SECTION 3

Bibliometric/scientometric review

Section overview

What is a bibliometric/scientometric review?

Sai Zhang & Vahid Aryadoust
Hebei Normal University | Nanyang Technological University

Definition

A bibliometric/scientometric review is a type of research synthesis that "encompasses all quantitative aspects and models related to the production and dissemination of scientific and technological knowledge" (De Bellis, 2009, p. 3). It is valuable for identifying influential research, revealing developmental trends, and evaluating scholarly impact within a particular academic domain. This section does not make a clear-cut distinction between bibliometrics and scientometrics due to their considerable overlaps, despite that the former deals with more general information about publications and the latter focuses on scientific communication (Aryadoust, 2023).

Methodological stages

Below are the major stages for conducting a bibliometric/scientometric review:

1. Study design

As in every research, the first stage of bibliometric/scientometric reviews begins with the study design. Scholars clearly define the research aim(s), specify the research question(s) that best reflect the aim(s), and choose the appropriate bibliometric/scientometric approach that addresses the question(s).

2. Data collection

Data collection is mainly divided into four sub-stages. It starts with choosing a bibliographic database (e.g., Web of Science, Scopus, and Google Scholar) that provides comprehensive literature coverage of the studied field and reliable sources of data. With the selected database, scholars define search terms and apply filters with a set of inclusion and exclusion criteria to ensure that the sample of

https://doi.org/10.1075/rmal.17.05zha

publications aligns with the study aim(s). The next sub-stage is to retrieve all the required data from the database and convert them into a suitable format for the applied bibliometric tool. Data cleaning is the final step, which often involves tasks like correcting errors, handling missing information, removing duplicates, etc. To enhance the transparency and comprehensiveness of the data collection process, researchers should draw upon a systematic review protocol, such as the PRISMA statement (Page et al., 2021) or the SPAR-4-SLR protocol (Paul et al., 2021), and offer scientific justifications for the decisions made at every sub-stage.

3. Data analysis

There are two main analysis procedures in bibliometric/scientometric reviews: performance analysis and science mapping. Scholars carry out performance analysis to evaluate the contributions of scientific actors based on a variety of bibliometric indicators, so the general scholarly landscape of the studied field can be presented. Science mapping, which involves tasks like co-citation analysis, co-word analysis, and bibliographic coupling, is performed to reveal the interactions between scientific actors.

4. Data visualization

This stage is to create graphical representations of the analysis, which makes the results easier to interpret and understand. Specific software programs can be applied to directly display the visuals, particularly in science mapping studies.

5. Interpretation and reporting

The final stage is to interpret the results, acknowledge any limitations, and suggest future directions. Also, it is of great importance to discuss how well the study supports the intended research aim(s) and addresses the research question(s).

Software/resources

Several software tools are available for conducting bibliometric/scientometric reviews, including the following options.

1. Bibliometrix (https://www.bibliometrix.org)

Bibliometrix is an open-source R package that offers a comprehensive set of tools for bibliometric/scientometric analyses. Its great flexibility and functionality make it a popular choice among researchers.

2. Biblioshiny (https://www.bibliometrix.org/biblioshiny)

Biblioshiny is a web-based application powered by the Bibliometrix R-package. Its user-friendly and well-structured interface allows researchers to perform in-depth analyses and visualizations without the knowledge of coding.

3. VOSviewer (https://www.vosviewer.com)

VOSviewer is a versatile tool for constructing and visualizing bibliometric networks, which represent the relationship between numerous elements of the scholarly world, such as journals, researchers, or publications. Additionally, this software is especially useful for analyzing textual data based on its text-mining functionality.

4. CiteSpace (http://cluster.cis.drexel.edu/~cchen/citespace)

CiteSpace is a Java-based application that helps analyze and visualize scientific patterns and developmental trends in a knowledge domain. It is perfect for researchers to track the research hotspots and fronts in any given field or discipline.

5. BibExcel (https://homepage.univie.ac.at/juan.gorraiz/bibexcel)

BibExcel is designed for preprocessing and analyzing bibliographic data. While it lacks built-in visualization features, it supports various export options that enable visualization through other software.

AI use declaration statement

The authors confirm that no AI tools were used in the preparation, writing, formatting, and editing of this work.

Examples of bibliometric/scientometrics publications

Here are four publications that are good examples of conducting this kind of research synthesis:

Meihami, H., & Esfandiari, R. (Ed.). (2024). *A scientometrics research perspective in applied linguistics*. Springer.

This book provides a comprehensive introduction to scientometrics in advancing and reshaping the field of applied linguistics. Through a blend of theoretical insights and hands-on guidance, it empowers readers to engage critically with scholarly production and communication using scientometric methods and tools.

Plonsky, L. (2023). Introducing bibliometrics in applied linguistics. *Studies in Second Language Learning and Teaching*, *13*(4), 721–726.

The editorial introduces a special issue of *SSLLT* on bibliometrics in applied linguistics. It highlights that bibliometric research is a natural extension of ongoing methodological innovations in the field, equipped with both methodological toolkits and synthetic mindedness.

Zakaria, A., & Aryadoust, V. (2024). A scientometric analysis of applied linguistics research (1970–2022): Methodology and future directions. *Applied Linguistics Review*, *15*(4), 1595–1619.

This article provides a systematic mapping of the applied linguistics field over five decades. With a clear how-to-do guide, it is a valuable reference for researchers interested in conducting bibliometric/ scientometric reviews.

Zhang, S., & Aryadoust, V. (2025). Why are some articles highly cited in applied linguistics? A bibliometric study. *Studies in Second Language Acquisition*, 1–23.

This bibliometric review brings a nuanced perspective on the dynamics of citation practices in applied linguistics research. Focusing on highly cited papers and underlying mechanisms that contribute to their prominence, it offers valuable insights for evaluating research performance and impact in the field.

References

Aryadoust, V. (2023). Bibliometrics and scientometrics in applied linguistics: Epilogue to the special issue. *Studies in Second Language Learning and Teaching*, *13*(4), 925–933.

De Bellis, N. (2009). *Bibliometrics and citation analysis: From the science citation index to cybermetrics*. The Scarecrow Press.

Page, M., McKenzie, J. E., Bossuyt, P. M., & Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *Systematic Reviews*, *10*, 89.

Paul, J., Lim, W. M., O'Cass, A., Hao, A. W., & Bresciani, S. (2021). Scientific procedures and rationales for systematic literature reviews (SPAR-4-SLR). *International Journal of Consumer Studies*, *45*(4), O1–O16.

CHAPTER 5

Methodological considerations in conducting bibliometric/scientometric reviews

Peter Crosthwaite, Sulistya Ningrum & Jihua Dong
University of Queensland | State Polytechnic of Sriwijaya | Shandong University

Bibliometric and scientometric reviews have become vital in Applied Linguistics, leveraging accessible data from repositories like Scopus and Web of Science to systematically map research outputs. These reviews quantitatively analyse scholarly publications to uncover research topic evolution, key authors, influential articles, citation patterns, and collaboration networks. Scientometric analysis extends this by exploring connections between science and societal, technological, and economic factors. With the surge in scholarly outputs in Applied Linguistics over the past two decades, robust methodological considerations are essential for reliable and transparent bibliometric reviews. This chapter examines these considerations, drawing on two representative studies by the authors focusing on corpus linguistics and written corrective feedback. It discusses critical decisions covering expected (i.e. planned for) and unexpected (i.e. resulting from the data collection and analysis) issues including keyword selection for data retrieval, inclusion/exclusion criteria, challenges with bibliometric data, software and statistical choices, and ensuring diversity and equity in reporting. By synthesising insights from these studies and others, this chapter offers a comprehensive guide for researchers, particularly novices, for conducting bibliometric reviews in Applied Linguistics.

Keywords: bibliometric review, scientometric review, corpus linguistics, metadata, applied linguistics

Background

With the advent of online research data repositories such as Scopus and Web of Science allowing ease of access to bibliometric data, bibliometric and scientometric reviews have gained importance in many disciplines — including Applied Lin-

https://doi.org/10.1075/rmal.17.05acro
Available under the CC BY-NC-ND 4.0 license.© 2026 John Benjamins Publishing Company

guistics — as systematic methods to map and analyse the research outputs of a given field. By quantitatively analysing scholarly publications through bibliometric analysis, these reviews reveal the evolution of research topics, key contributing authors, the most influential articles or venues within a specific area, citation patterns, and institutional collaboration networks. Unlike meta-analysis (Chapters 7, 8), bibliometric methods typically do not estimate effect sizes but instead examine structural patterns in the literature. The scientometric analysis goes one step further, using bibliometric approaches to analyse the relationships between science and societal, technological, and economic factors using data sources including not only bibliometric data but patent analysis and research funding, although such analyses are as yet uncommon in Applied Linguistics (see Zakaria & Aryadoust, 2024 however for a notable exception).

With marked increases in scholarly outputs in Applied Linguistics over the past two decades (Lei and Liu, 2019), the need for robust methodological considerations in conducting such reviews is crucial to ensure reliability, transparency, and the representativeness of any findings (see Chapter 1). This chapter therefore explores the methodological considerations involved in conducting bibliometric reviews with insights drawn from two representative studies in Applied Linguistics published by the authors of this chapter, specifically focusing on corpus linguistics and written corrective feedback (WCF). We discuss the key methodological decisions made when conducting these reviews, including the selection of appropriate keywords for data retrieval, the study's inclusion and exclusion criteria for retrieved studies, the challenges of working with bibliometric data, the choice of software and statistical approaches for analysis, and diversity and equity in reporting results. Drawing on our experiences with these representative bibliometric studies and other studies conducted by the third author and other researchers, we aim to provide a comprehensive guide for researchers, especially novices, who may consider conducting their own bibliometric reviews in Applied Linguistics.

1.1 Representative study 1: A bibliometric review of corpus linguistics (Crosthwaite, Ningrum & Schweinberger, 2022)

Crosthwaite, P., Ningrum, S., & Schweinberger, M. (2022). Research trends in corpus linguistics: A bibliometric analysis of two decades of Scopus-indexed corpus linguistics research in arts and humanities. *International Journal of Corpus Linguistics, 28*(3), 344–377. https://doi.org/10.1075/ijcl.21072.cro

The first representative study in this chapter presents a bibliometric analysis of the field of corpus linguistics. Corpus linguistics involves the principled collection of electronically searchable texts aiming to represent the genre, register or

language under investigation, before analysis using specialised software (known as corpus tools) that can retrieve textual, statistical and visual data revealing the patterns of language use contained within. This information can be used to describe, compare, and teach language trends used for a wide range of linguistic and pedagogical purposes. The bibliometric analysis of corpus linguistics by Crosthwaite, Ningrum, and Schweinberger (2022) mapped the important authors, publications, and regions of corpus linguistics research in the arts and humanities between 2000 and 2020. The study included 5,829 articles from 429 Scopus-indexed journals that met the inclusion criteria for the study. The data revealed an increase in research on corpus-assisted discourse studies, lexical bundles, and academic writing over more established research areas, such as grammar, alongside the formation of newer topics such as multilingualism and social media. The study also highlighted the rise of China in corpus linguistics research, reflecting its increasing importance as a site of Applied Linguistics research more broadly, as well as an increase in corpus studies on Chinese, Russian, Spanish, and Italian languages. The most influential contributors to the field in terms of citation and publication counts as well as the most influential individual works were also revealed. The results indicated multiple connections between advancements in corpus research and historical socio-cultural changes in Applied Linguistics and society at large.

1.2 Representative study 2: A bibliometric review of L2 written corrective feedback

Crosthwaite, P., Ningrum, S., & Lee, I. (2022). Research trends in L2 written corrective feedback: A bibliometric analysis of three decades of Scopus-indexed research on L2 WCF. *Journal of Second Language Writing, 58*, 100934. https://doi.org/10.1016/j.jslw.2022.100934

WCF is a key pedagogical approach in the field of second language writing teaching and learning involving teachers' marking of errors in learners' written production to facilitate subsequent revision for remediation of said errors, in order to improve both writing quality and promote conditions for second language acquisition. In our bibliometric study, Crosthwaite, Ningrum, and Lee (2022) mapped 30 years of Scopus-indexed research on L2 WCF, analysing 493 journal articles, books and book chapters meeting the study's inclusion criteria. We tracked key research trends such as teacher and peer feedback, direct versus indirect WCF, learner engagement with WCF, and the role of WCF in improving grammatical accuracy. Notably, the data reveal the field has witnessed significant geographical diversification, with an increasing number of studies emerging from China, Iran, and Spain, moving beyond an early dominance of studies from the

USA and UK. However, studies on languages other than English remain infrequent. A number of influential, highly cited L2 WCF researchers were identified in the data over the study period, including clusters of L2 WCF researchers spanning a range of (inter)disciplinary research areas.

2. Methodological and ethical issues

When planning and conducting a bibliometric review, researchers face various methodological and ethical dilemmas. These can be broadly categorised into anticipated and unanticipated issues, and both must be addressed thoughtfully to ensure the validity and reliability of the review.

2.1 Anticipated methodological and ethical issues

In conducting a bibliometric analysis, certain methodological and ethical decisions must be made by the researchers ahead of retrieving the metadata for subsequent analysis. These are **anticipated** in that appropriate retrieval of said metadata cannot be conducted without first considering these issues, which are mostly related to the inclusion and exclusion criteria for metadata retrieval. This includes choosing the indexing database from which to retrieve article metadata for analysis (e.g., Scopus, Web of Science), the selection of keywords for retrieval, the types of documents to include (e.g., journal articles, book chapters, conference papers), whether the publications are available open access or behind subscriptions, the timeframe(s) from which to include publications, and the ethical considerations underlying many of these choices.

The selection of indexing databases from which to retrieve article metadata for inclusion is of course crucial to the whole exercise. Typically, Scopus or Web of Science are used in bibliometric studies due to their broad coverage of journals, detailed metadata including publication date, author citations, author affiliations, co-authorship networks, and journal impact factors. They also allow for ease of retrieval and download of said metadata for subsequent analysis. In the two representative studies of this chapter, Scopus allowed for the download of metadata across journal articles, books, and book chapters using several simple checkboxes and sliders following keyword input, downloading these into a .csv file for subsequent analysis. However, in our L2 WCF study, for example, this approach omitted some publication outlets that are gaining in prominence only recently (e.g. Journal of Writing Research), or L2 writing-focused journals not indexed in Scopus (e.g. Journal of Response to Writing). We describe studies using alternatives to Scopus in **Section 4** of this chapter. Another alternative is to select only publications from certain journals (Chapters 2, 3).

Another crucial factor is the **type of publications** to be retrieved for analysis. These may include journal articles, academic books, book chapters, conference papers, undergraduate, graduate or PhD level theses/dissertations, pre-prints, websites and many more. The decision on what to select will depend on the relative level of peer review and study quality one may wish to capture in their bibliometric study. In our L2 WCF study, only final-stage journal articles, books, and book chapters were included. This decision, while ensuring high-quality peer-reviewed publications, may have omitted significant yet unpublished research in conference papers for example, which in certain fields (e.g. computational linguistics) may be more prestigious than publication in journal articles. An even narrower approach was taken in our corpus linguistics study where only final-stage journal articles were included given the vast size of the original dataset, therefore omitting influential book chapters and monographs (although the latter were discussed in terms of which works were cited in our dataset of research articles).

Another anticipated issue is **choosing open- versus closed-access publications**. Open-access sources often have higher citation rates due to their broader availability, which can skew results. Researchers must consider whether to include only open-access journals or to balance their dataset with closed-access ones, acknowledging that the accessibility of research can affect its inclusion and impact in bibliometric reviews (see Zhang & Aryadoust, 2025 for evidence of this).

It is also important to consider the **selection of (sub)disciplines for analysis**. This can include the selection of discipline when retrieving articles from the repository itself in the first instance (if the repository allows for this), or it can come in the form of selecting journals for inclusion from a broader disciplinary category (e.g. arts and humanities; social sciences). For example, in our bibliometric study of corpus linguistics, we limited the scope of publications to those of the arts and humanities, although this decision resulted in papers from computational linguistics, which often report corpus-based research, being omitted from inclusion. Journals in this subdiscipline often fall under the sciences and tend to focus on the tool or program accuracy over the detailed analysis of corpora. Papers that span multiple fields may also not be fully captured or appropriately classified. Researchers considering conducting a bibliometric study should therefore have a good general knowledge of their field so as to understand which journals fall under their specific discipline and which do not, as retrieval of metadata from repositories such as Scopus may result in journals not strictly within the target discipline being included.

Another key consideration is the **timeline(s) for inclusion.** Bibliometric studies on relatively new topics (e.g. generative AI) will result in a relatively narrower timeline from which to draw related research, meaning that citation counts for authors and articles will be skewed towards those generating high immediate

impact but potentially without long-lasting impact. There may also be insufficient citation counts for subsequent analyses if the timeline is too narrow. Moreover, studies that seek to compare bibliometric trends over set periods such as decades (as conducted in both our representative bibliometric studies in this chapter) run the risk of earlier decades having relatively little in the way of publications, given the far smaller number of journals available decades ago compared to the vast array of journals now available in our field. For example, in our L2 WCF study, the decision was made to retrieve data from 1991 onwards, as prior to 1991 only four representative studies met our inclusion/exclusion criteria.

Defining the research area and **selecting appropriate** keywords are also critical to the success of any bibliometric analysis. Keywords must reflect the breadth of the field while ensuring sufficient specificity to capture all relevant publications, so simply searching for a single term may be insufficient. For instance, in WCF research, terms like "written feedback," "corrective feedback," and "error correction" are frequent in L2 writing research, but each carries slightly different connotations, potentially leading to the inclusion or exclusion of studies that do not fit the intended focus. For instance, studies on general written feedback or those on *oral* corrective feedback are not to be mistaken for studies on WCF. Likewise, studies that do not specify an L2 focus or with unclear L2 population need to be excluded. Additionally, in our study on corpus linguistics, what counts as a 'corpus' can vary significantly across disciplines. For example, 'corpus' is used in a non-linguistic sense (e.g., in art, referring to a body of work, or in archaeology, referring to the collection of retrieved artefacts from a particular location). Moreover, as highlighted in Dong et al. (2023a), the very definition of the target construct for analysis (in their case, 'data-driven learning') may have evolved over time, from its conceptualisation in the 1980s to the modern day. This temporal evolution presents a methodological challenge for bibliometric studies, as key terms and conceptual boundaries shift in response to theoretical advancements, technological developments, and pedagogical refinements. To address such issue, researchers may employ a diachronic perspective in bibliometric analyses by segmenting their dataset into temporal phases, allowing for a more fine-grained analysis of how a field has developed across distinct scholarly periods, as illustrated in Dong et al. (2024). Such an approach enables scholars to distinguish between foundational works that established the field, intermediary studies that expanded its theoretical scope, and contemporary research that reflects its latest innovations.

The choice of **bibliometric software** is another anticipated issue. In our representative studies, VOSviewer was used for its ability to visualise co-citation networks and keyword co-occurrence maps, making it ideal for identifying research trends. Additionally, VOSviewer's hierarchical clustering algorithms allowed for

the creation of clear, visually interpretable networks, although it does not have the time-slicing feature that allows users to conduct historical analyses to see how the influence of certain research topics has evolved, useful for identifying long-term shifts in research trends (Zhang & Chen, 2020). Other tools are discussed in **Section 4**.

The **selection of quantitative metadata for analysis** is also a major methodological decision to be considered. For example, do we compare author impact through the number of all citations over time (thus skewing the data towards older, established researchers), or citations in a certain period? Measuring and comparing author productivity is also difficult when comparing authors who mostly publish books versus those who publish journal articles, or those who produce non-traditional research outputs that could still be considered Applied Linguistics e.g. dictionaries, reference guides, and performances. Other researchers may be very productive in the field but choose to publish in outlets that are not indexed in research repositories (e.g. Scopus), whose work, therefore, will be omitted from subsequent analysis.

Conducting bibliometric analyses raises several ethical concerns. While ethical approval from university IRB boards is not typically required for Bibliometric studies as no human data is required (typically ethics waivers may be arranged, as they were for the two studies mentioned in this chapter), ethics in a broader sense is fundamental to any proposed bibliometric study's methods (Chapter 20). It is important to ensure objectivity and systematicity while avoiding biases during dataset construction and subsequent analysis. For example, limitations in metadata inclusion, such as whether publications are limited to English, may reduce dataset reliability or promote marginalisation of researchers based on socio-economic background, gender, or institutional status. Another significant issue involves the inclusion of publications from predatory journals, which may engage in unethical practices like citation manipulation or publishing fabricated research. Nowadays, there is a growing presence of predatory journals in mainstream databases, requiring researchers to have a degree of expertise in the relevant discipline to be able to spot predatory journals within the sample dataset. This can be achieved by checking for suspicious signs such as unrealistic publication timelines between submission and acceptance as shown on a journal's website, by consulting experienced mentors or using resources like Beall's List[1] to verify journal credibility. Cross-referencing a journal's indexing in trusted databases and scrutinising its editorial board for recognised experts can further help distinguish legitimate publications from predatory ones.

1. https://beallslist.net/

2.2 Unanticipated methodological and ethical issues

In contrast to the expected methodological and ethical issues in the previous section, a range of **unanticipated issues** arise typically following metadata retrieval from the target repository, during analysis, or even following the publication of the study itself. This section discusses several of these issues.

One key limitation in retrieving datasets from Scopus is the **restriction on exporting metadata** for a large number of documents. Scopus allows users to export metadata using various filters; however, the maximum limit for exporting records is 20,000 per seven days. These records can be exported in formats such as CSV, TXT, BIB, or RIS. As our dataset exceeded this limit, we had to split the data into smaller batches and download them separately, which can be time-consuming and error prone. This limitation becomes particularly problematic for bibliometric studies requiring comprehensive datasets involving tens of thousands of records. At the time our articles were written, the Scopus API would send an email with a link to download the CSV dataset, which was accessible for one week. This process often introduced delays as we had to wait for a new download link to be sent, and also raised difficulties if we needed to re-download a list we had previously downloaded. Additionally, managing and merging multiple exported files to form a complete dataset can pose challenges in ensuring data consistency and avoiding duplication.

Unanticipated issues often arise with **missing data**. For example, in our representative studies, missing abstracts or keywords in many documents indexed in Scopus made it difficult to determine their relevance without accessing the full text. This can lead to the exclusion of potentially important studies because of incomplete metadata (particularly if the researchers lack access to the work(s) in question), skewing the dataset and influencing the resulting analysis. For example, in our L2 WCF study, a number of articles were excluded where some form of WCF might well have been employed (e.g., peer feedback during online collaborative writing) but they were not specifically analysed. Another factor is that the chosen repository (e.g., Scopus) may not be able to provide information on how often articles were cited within a given timeframe (e.g. how many citations were generated between 2001–2005, 2006–2010, etc.), meaning the number of citations would only reflect the most recent citation count the repository holds. Article keywords were also largely absent from articles in our L2 WCF study between 1991 and 2000 as this timeframe dated back to pre-broadband internet and the wide accessibility of journals online. Studies on languages other than English may also be rare in certain fields, making subsequent analysis tricky. For example, in our L2 WCF study, only one study in a language other than English (Spanish) made our inclu-

sion/exclusion criteria between 1991–2000. This meant this time frame could not then be included in a subsequent regression analysis due to a lack of data.

Duplicate records are also an unexpected issue that may need to be resolved. For example, if a given study has been published both as a conference paper and a journal article with the same title, a decision should be made as to whether both entries should be combined for subsequent analysis. Similarly, when analysing author metrics, consideration for authors' initials is important in the case where multiple entries are found for the same person, e.g. M. Halliday vs. M.A.K. Halliday. Sometimes, researchers may face the opposite problem where authors with the same name and initial (e.g. Y. Kim) may appear to be the same person but are in fact different authors. **Inappropriate records** also appear in datasets retrieved from repositories such as Scopus if they have been incorrectly tagged. For example, spelling errors in journal, title, or author names may require significant manual correction, while non-research focused articles such as editorials, corrigendums, retraction notices, letters to the editors, and commentaries may be included in the initial data retrieval but need to be manually excluded later.

Another limitation of bibliometric analysis is the challenge of **time lag** in the rapidly evolving academic landscape. Bibliometric studies rely primarily on published references, meaning that their findings inherently reflect past research trends rather than capturing real-time developments. As a result, there may be an unavoidable time lag between the latest advancements in a field and their visibility in bibliometric analyses. This issue is particularly notable in fast-moving areas of Applied Linguistics, where emerging theoretical frameworks, methodological innovations, and technological advancements may not yet be adequately represented in citation databases. Additionally, the indexing and updating processes of bibliometric databases such as Scopus and Web of Science introduce further delays, as newly published works take time to be processed and integrated. Consequently, bibliometric analyses may not fully account for the most recent shifts in scholarly discourse, which can impact the comprehensiveness and relevance of their conclusions.

Researchers may also face **reactions from scholars** who feel that their work was unfairly excluded from a review. For them, this exclusion may lead to interpersonal and reputational challenges. In our own experience, some prominent researchers expressed concerns about not being included in a bibliometric study due to our predefined inclusion and exclusion criteria, even though their work was highly relevant to the field. This issue necessitates careful explanation regarding inclusion and exclusion criteria in the study's methodology section.

3. Course of action

When dealing with both expected and unexpected challenges, researchers must adopt flexible and transparent strategies. This section describes how the two representative studies listed in this chapter addressed these challenges.

To address the dataset export limitations in Scopus, we divided data into smaller subsets by applying filters such as publication year, subject area, or document type, and exported them in manageable batches. This approach ensured that the dataset remained within the 20,000-record export limit imposed by Scopus. Although the recent feature allowing users to directly download CSV files without relying on email has improved accessibility, it does not resolve the fundamental challenge of handling large datasets for bibliometric studies. To mitigate these challenges, employing robust data cleaning and preprocessing techniques becomes essential. We applied preprocessing techniques such as harmonising keywords, resolving variations in journal titles, and verifying publication dates to further enhance the accuracy and usability of the dataset.

Regarding the study's inclusion and exclusion criteria, a crucial factor in determining whether our criteria were appropriate was to benchmark our final dataset sample against that of another (smaller) sample that could also be considered representative of the area under investigation. For example, in the representative study on corpus linguistics, we compared the number of articles in our dataset with a smaller subset of articles taken from four key corpus linguistics-focused journals (*International Journal of Corpus Linguistics, Corpus Linguistics and Linguistic Theory, Corpora*, and *International Journal of Learner Corpus Research*). This was because if one can consider all articles from these journals as 'corpus linguistics research', the proportion of articles in our sample dataset should be as close to 100% of the list of all articles in these four journals as possible. Our final sample was determined to contain 555 out of a maximum 601 (92.35%) articles from these four journals within the specified timeframe, which, together with manual checking of each article in the sample, confirms our inclusion and exclusion criteria cover at least 90% coverage of corpus linguistics research.

A key mechanism for screening articles is ensuring interrater/intercoder reliability. In bibliometric analyses where large datasets are coded and categorised by multiple researchers, interrater/intercoder reliability is critical to ensure the validity of the results. In the studies by Dong et al. (2023a; 2023b), the researchers used Cohen's kappa to assess agreement in identifying discrepancies in data interpretation and ensuring consistency across reviewers. In the two representative studies in this chapter, each article was manually checked for exclusion and inclusion by two researchers using a colour coding scheme in two matching Excel tabs (green=agree, yellow=unsure, red=remove). Articles that were deemed irrelevant

by both researchers were excluded, leading to the creation of a new dataset for another round of colour coding. After five full rounds, any remaining yellow-coloured articles were discussed by the researchers before reaching a final agreement, ensuring both researchers reached 100% agreement on the final dataset.

Regarding unexpected issues related to duplicates in retrieved metadata, one solution taken in the two representative studies was to keep a thesaurus of article keywords which combines all variants of a keyword into a single unit. For example, in the corpus linguistics study, we merged the keywords 'CADS', 'CADA', 'corpus-assisted discourse study', 'corpus-assisted discourse analysis', and 'corpus-based discourse study', recording each as 'CADS'. Similarly, for languages, varieties such as 'Hong Kong English' and 'Indian English' were recorded as 'English'; 'Slovene' was merged with 'Slovenian', etc.

Regarding statistical analyses, a number of possibilities are available in terms of the selection of inferential statistical techniques, which are informed by the research questions the study is set to tackle. In the two representative studies in this chapter, we adopted regression analyses (using the R statistical package) to determine relative increases and decreases in topic popularity, languages of investigation, and geographic sources of publication. An additional step in ensuring the validity of our study in terms of external scrutiny was to hyperlink the full R notebooks used to conduct these regression analyses within the articles themselves, with the dataset also stored online for download. Another statistical course of action taken in the two representative studies was to normalise the frequency of keywords by timeframe to a value of *n* per 100. This was important in meaningfully comparing the relative increase/decrease of keywords within and across each timeframe, being mindful that earlier time frames tend to have a smaller dataset than more recent timeframes due to the ever-increasing volume of research publications over time. A similar approach was taken to analysing change over time in countries of publication origin, with a proportional percentage increase over time calculated using the geometric means of the earliest and latest timeframes. Regarding author citations, it can sometimes be the case that a single multi-authored article can be very highly cited. If the authors did not publish other cited papers in the dataset, when generating the tables containing the highest-cited authors it can appear as if a number of people have exactly the same number of citations. Therefore, in our representative studies, for a particular timeframe, we added a footnote to the tables informing the reader that authors with an asterisk (*) next to their names worked on the same article. Additionally, because Scopus could not provide the number of citations per individual timeframe, our citation counts only matched the current historical citation count at the time of metadata download. While this privileges older works in terms of citation count per timeframe, it at least suggests the authors featured in top-cited lists have published studies with a lasting legacy in that research area.

Regarding tensions from researchers in the field post-publication, it is important to thoroughly communicate inclusion and exclusion criteria both in the article itself as well as externally. In the review, this should be clear in the study's methodology section but also in the study's limitations section where all expected and unexpected issues are clearly acknowledged. Regarding external communication, in the two representative studies, we consulted with experts in the related field through personal correspondence and Applied Linguistics-related social media groups; we also sought feedback by presenting preliminary findings at conferences to gain the broader academic community's input. This open dialogue helped address concerns about the applied criteria and ensured that our dataset was as comprehensive as possible. As mentioned, making the sample datasets and R notebooks available through open-access methods may also increase transparency and help mitigate misunderstandings.

Regarding the ethical issues mentioned in Section 2, the key is to ensure that each of the methodological steps discussed above is conducted objectively and systematically without cherry-picking article types, sources, statistics, or coverage to ensure the data fit the researchers' pre-defined narrative(s). Moreover, clearly acknowledging the limitations of the researchers' approach to metadata inclusion is vital in ensuring the reliability of the final dataset including specifying if the retrieved articles included English-only publications or those in other languages. It is also important (particularly if the study is eventually published) to ensure that the selection of authors, institutions or regions for inclusion is, in fact, representative of the field, so as to avoid marginalising individual researchers or groups of researchers based on factors such as socio-economic background, gender, institution status. Finally, researchers need to consider how to handle publications from predatory journals. Such journals may engage in unethical research practices such as citation inflation/manipulation, publish research with false data or even articles entirely produced using generative AI. In the representative studies, this issue was addressed by manually reviewing the journals' quality (as discussed earlier in this chapter) to ensure only legitimate publications are featured, although expertise in the area may be a requirement for this.

4. Alternative courses of action

The chosen courses of action reflect those undertaken in the two representative studies outlined in this chapter. However, other options are available to researchers in overcoming both expected and unexpected issues that may arise when conducting a bibliometric analysis. This section describes how other

notable bibliometric studies in Applied Linguistics have overcome such issues for the reader.

Regarding expected issues relating to inclusion of articles for analysis, while our representative studies used Scopus, alternative courses of action could include using different indexing databases such as Web of Science, which offers better coverage in some areas of the humanities compared to Scopus. Recent Bibliometric studies in Applied Linguistics looking at the use of corpora for data-driven learning and metadiscourse for example (Dong et al., 2023a, 2023b, 2024) used Web of Science, which contains metadata on articles dating back much further than Scopus and has stricter quality criteria for publications to be included in its database, albeit at the expense of the coverage offered by Scopus. Additionally, Web of Science's Arts & Humanities Citation Index (A&HCI) is a specific feature that provides superior coverage in these disciplines (Mongeon & Paul-Hus, 2016). Web of Science also collects citations on software and datasets, an important omission in our Scopus-driven bibliometric study on corpus linguistics as many top-cited researchers in the field generate 1000s of direct citations to their corpus analysis software or corpus data rather than papers associated with said software or data. In addition, Web of Science combines early access and published works into a single record, expanding the potential range of articles for analysis. Another strategy is expanding database coverage by utilising the freely accessible Google Scholar which indexes the full text or metadata of literature across a variety of disciplines. Although time-consuming, this methodology could have expanded the dataset to include grey literature and non-indexed works. Some studies have used Google Scholar for this purpose (e.g., Bornmann et al., 2016). However, despite having the largest coverage, third-party software must be used to extract article metadata and Google Scholar cannot account for low-quality or predatory publications. Regional databases such as SciELO for publications in Latin America, Spain, and Portugal or the China National Knowledge Infrastructure may also be used to gather more targeted data.

Regarding the bibliographic analysis software, while our representative studies used VOSViewer, other tools offer alternative functionalities and may be more appropriate depending on the goals of the review. Tools such as OpenRefine, Python, and specialised bibliometric software can be highly effective in consolidating multiple files into a cohesive dataset (Moral-Munoz, et al., 2019). OpenRefine excels in cleaning and transforming data, making it ideal for identifying and resolving inconsistencies. Python, with its extensive libraries like pandas and NumPy, offers powerful scripting capabilities for automating complex data integration processes. Meanwhile, bibliometric software like VOSViewer,[2] BibEx-

2. https://www.vosviewer.com/

cel,[3] CiteSpace, or Gephi provides solutions for managing and analysing publication data, ensuring seamless integration of metadata. By leveraging these tools, researchers can streamline merging of datasets while maintaining accuracy and consistency (Moral-Munoz et al., 2019). A comprehensive overview of all available tools for analysis is found in Zakaria & Aryadoust (2024). For example, Gephi[4] offers deeper insights into citation clusters and co-authorship networks. Meanwhile, CiteSpace[5] excels at identifying upcoming trends and citation hotspots. It combines information visualisation, data mining, and bibliometrics to automatically identify clusters and emphasise temporal trends in a study subject (Chen, 2018; Liu & Wang, 2016). Citespace was used in Dong et al.'s (2023a) bibliometric analysis of metadiscourse as an evolving area of Applied Linguistics research as this facilitates co-citation analysis. This involves developing a dynamic co-citation network comprising references co-cited in all studies within the dataset. This allows researchers to determine the themes of each co-citation cluster (and therefore, the target research area) by extracting noun phrases from the titles, keywords, or abstracts of articles that cited the respective clusters (Chen, 2016).

Regarding the choice of statistics for analysis, Dong et al. (2023b) used bibliometric techniques to explore the field of data-driven learning, a sub-field of corpus linguistics specifically dedicated to the use of corpora for language teaching and learning. The study uses Shneider's (2009) four-stage evolutionary model of a research field, divided into a 'conceptualising stage' (1980s–1998), the 'maturing stage' (1998–2011), and the 'expansion stage' (2011–now). Their study used keyword co-occurrence analysis of article keywords to determine the top keywords over a thirty-year publication period, and co-citation analysis of influential publications and journals. Importantly, their study also uses Structural Variation Analysis (SVA), an alternative statistical approach to (co)citation tracking, that charts 'transformative' individual research outputs responsible for influencing a field during these evolutionary periods. 'Transformative' in this sense can be quantified through SVA based on information in the publication, mainly cited references. The higher the degree of variation, the more transformative a publication will be. This approach intends to negate the impact of chronological bias inherent in (co)citation practices where newer articles receive fewer citations than older articles, serving as "a good indicator of potentially high-impact publications (regardless of their publication dates), and thereby signalling the direction of future research in a field" (p. 342). SVA has been used to identify studies with high transformative potential such as Nobel Prize-winning publications (Sebastian &

3. https://homepage.univie.ac.at/juan.gorraiz/bibexcel/

4. https://gephi.org/

5. https://citespace.podia.com/

Chen, 2021). This approach was also used in Dong et al.'s (2024) bibliometric analysis of second language writing assessment, identifying 11 recent publications with 'transformative potential' as realised through SVA between 2018 and 2022. Given the advantages of SVAs in mitigating chronological bias and capturing emerging high-impact research, its application in Applied Linguistics bibliometric studies holds significant promise. Future research could extend its use to other subfields, such as multilingualism, language assessment, and technology-enhanced language learning, to uncover evolving paradigms and the scholars shaping these developments. Furthermore, integrating structural variation analysis with network-based visualizations, such as bibliographic coupling and thematic clustering could provide a more comprehensive understanding of how knowledge structures in Applied Linguistics evolve over time, offering valuable insights for both researchers and policymakers in the field.

Another interesting approach is found in a bibliometric study exploring why some articles become highly cited in Applied Linguistics (Zhang & Aryadoust, 2025), who applied multiple linear regression analysis techniques to identify 11 article-extrinsic factors predictive of highly cited articles, including number of authors, subfield, methodology, title length, CiteScore, accessibility, and scholar h-index. According to the study's authors, such statistical analysis allows researchers to "recognize the underlying forces affecting research impact" (p. 1).

To address the limitation of outdatedness in bibliometric analysis, researchers can implement several strategies to improve the timeliness and relevance of their findings. One approach is to incorporate preprints and grey literature, such as conference proceedings and working papers, which can capture emerging research before it is indexed in bibliometric databases. Additionally, adopting a longitudinal approach by periodically updating datasets enables researchers to track evolving research trends more accurately. Integrating bibliometric analysis with expert review or qualitative methods can further contextualize findings, offering insights into recent developments that may not yet be reflected in citation networks. Expanding database coverage by cross-referencing multiple sources, such as Scopus, Web of Science, and Google Scholar, as mentioned previously can help mitigate discrepancies in indexing speed and ensure more comprehensive data collection. Finally, explicitly acknowledging the inherent time lag in bibliometric studies allows readers to interpret findings with an awareness of potential gaps, while also encouraging continuous updates to maintain the accuracy and relevance of bibliometric analyses as the field progresses.

5. Conclusion

Bibliometric and scientometric reviews are powerful tools for mapping research landscapes, but they come with significant methodological and ethical challenges. By carefully selecting inclusion criteria, defining constructs, and choosing appropriate software, researchers can produce meaningful and equitable reviews. Being transparent about the decision-making process and open to feedback from the academic community helps ensure that the results are robust and accepted by the wider field. We hope the Applied Linguistics community can benefit from the experiences outlined in this chapter, and look forward to more bibliometric studies of key issues in Applied Linguistics in the near future. It is also hoped that this chapter will serve as a foundation for further exploration and innovation in bibliometric studies of Applied Linguistics, fostering a more nuanced and comprehensive understanding of a particular intellectual landscape.

6. Questions for discussion

1. How can researchers ensure that bibliometric reviews account for diversity and equity in terms of geographical representation?
2. What are the potential risks of relying solely on Scopus or Web of Science for bibliometric analyses? What other databases can be used?
3. What are the limitations of bibliometric software in revealing emerging trends in certain fields?
4. What ethical considerations arise when excluding certain types of publications (e.g., non-English language papers) from a bibliometric review?
5. How can researchers handle backlash from scholars whose work is excluded from a bibliometric review?

AI use declaration statement

During the preparation of this work, the authors used ChatGPT 4.0 for editing and proofreading the work, including reducing the word count. As authors, we confirm that after using the tool, we reviewed and edited the content as needed and take full responsibility for the content of the publication, in alignment with COPE and the publisher's AI policy.

Further reading

Lei, L., & Liu, D. (2019). Research trends in applied linguistics from 2005 to 2016: A bibliometric analysis and its implications. *Applied Linguistics*, *40*(3), 540–561.

This study is valuable for its comprehensive analysis of research trends in applied linguistics using bibliometric methods, offering insights into topic shifts and citation patterns that guide researchers in understanding the field's evolution.

Zakaria, A., & Aryadoust, V. (2024). A scientometric analysis of applied linguistics research (1970–2022): Methodology and future directions. *Applied Linguistics Review*, *15*(4), 1595–1619.

This work provides a robust methodological framework for scientometric analysis, making it an excellent resource for researchers seeking to map long-term trends and methodological advancements in applied linguistics.

Sun, Y., & Lan, G. (2021). Research trends in 'trans-'studies on writing: A bibliometric analysis. *System*, *103*, 102640.

Focused on translingual writing, this study demonstrates the application of bibliometric techniques to niche areas, offering practical guidance for analysing specific subfields within applied linguistics.

References

Bornmann, L., Thor, A., Marx, W., & Schier, H. (2016). The application of bibliometrics to research evaluation in the humanities and social sciences: An exploratory study using normalized Google Scholar data for the publications of a research institute. *Journal of the Association for Information Science and Technology*, *67*(11), 2778–2789.

Chen, C. (2016). *CiteSpace: A practical guide for mapping scientific literature*. Nova Science Pub.

Chen, C. (2018). A bibliometric analysis of research on the Internet of Things: A CiteSpace study. *Journal of Information Science*, *44*(6), 872–885.

Crosthwaite, P., Ningrum, S., & Lee, I. (2022). Research trends in L2 written corrective feedback: A bibliometric analysis of three decades of Scopus-indexed research on L2 WCF. *Journal of Second Language Writing*, *58*, 100934.

Crosthwaite, P., Ningrum, S., & Schweinberger, M. (2022). Research trends in corpus linguistics: A bibliometric analysis of two decades of Scopus-indexed corpus linguistics research in arts and humanities. *International Journal of Corpus Linguistics*, *28*(3), 344–377.

Dong, J., Dong, S., & Buckingham, L. (2023a). How does a research topic evolve into a research field? A bibliometric analysis of metadiscourse research. *Ibérica*, *45*, 163–189.

Dong, J., Zhao, Y., & Buckingham, L. (2023b). Charting the landscape of data-driven learning using a bibliometric analysis. *ReCALL*, *35*(3), 339–355.

Dong, J., Zhao, Y., & Buckingham, L. (2024). Thirty years of writing assessment: A bibliometric analysis of research trends and future directions. *Assessing Writing*, *61*, 100862.

Lei, L., & Liu, D. (2019). Research trends in applied linguistics from 2005 to 2016: A bibliometric analysis and its implications. *Applied Linguistics*, *40*(3), 540–561.

Liu, H., & Wang, X. (2016). A bibliometric analysis of the research on sustainable tourism: A network analysis approach. *Journal of Sustainable Tourism*, *24*(1), 1–22.

Mongeon, P., & Paul-Hus, A. (2016). The journal coverage of Web of Science and Scopus: A comparative analysis. *Scientometrics*, *106*(1), 213–228.

Moral-Munoz, J.A., López-Herrera, A.G., Herrera-Viedma, E., & Cobo, M.J. (2019). Science mapping analysis software tools: A review. In W. Glänzel, H.F. Moed, U. Schmoch, & M. Thelwall (Eds.), *Springer handbook of science and technology indicators* (pp. 159–185). Springer.

Sebastian, Y., & Chen, C. (2021). The boundary-spanning mechanisms of Nobel Prize winning papers. *PLOS ONE*, *16*(8), e0254744.

Shneider, A.M. (2009). Four stages of a scientific discipline; four types of scientist. *Trends in Biochemical Sciences*, *34*(5), 217–223.

Zakaria, A., & Aryadoust, V. (2024). A scientometric analysis of applied linguistics research (1970–2022): Methodology and future directions. *Applied Linguistics Review*, *15*(4), 1595–1619.

Zhang, S., & Aryadoust, V. (2025). Why are some articles highly cited in applied linguistics? A bibliometric study. *Studies in Second Language Acquisition*, 1–23.

Zhang, Y., & Chen, Y. (2020). Trends and hotspots of research on sustainable urbanization: A bibliometric analysis using CiteSpace. *Sustainability*, *12*(6), 5432.

CHAPTER 6

Towards methodological rigor and analytical depth in conducting bibliometric reviews

A case study in Applied Linguistics

Sai Zhang & Vahid Aryadoust
Hebei Normal University | Nanyang Technological University

This chapter critically reflects on our bibliometric review (Zhang & Aryadoust, 2025), which examined the role of extrinsic factors in shaping the citation landscape of Applied Linguistics literature. Accordingly, we highlight two key methodological considerations in conducting bibliometric reviews. The first purpose of this chapter is to explain our methodological approach and document the research process in a transparent and systematic manner so that the study can be understood, evaluated, and potentially replicated by other researchers. The second purpose of this chapter is to demonstrate our effort to identify potential biases and social dynamics underlying citation practices within the disciplinary context and to highlight the importance of connecting bibliometric findings to broader theoretical and practical implications. This chapter would be useful for researchers interested in conducting bibliometric reviews to get a deeper engagement with this synthetic approach in Applied Linguistics and beyond.

Keywords: applied linguistics, bibliometric reviews, citation practices, methodological framework

1. Introduction

This chapter examines key principles of bibliometric reviews as a basis for exploring how extrinsic factors influence the citation patterns within Applied Linguistics literature. In a recent bibliometric study (Zhang & Aryadoust, 2025), we focused on the ongoing debates over what is actually measured by citations. Despite the widespread reliance on citation counts as a proxy for scholarly merit, empiri-

https://doi.org/10.1075/rmal.17.06zha
Available under the CC BY-NC-ND 4.0 license.© 2026 John Benjamins Publishing Company

cal evidence across disciplines has demonstrated that equating citation frequency with research quality can be problematic (Kousha & Thelwall, 2024). We thus aimed to discuss the dynamics and social mechanisms that contribute to highly cited Applied Linguistics papers, as this would promote a more realistic and context-sensitive interpretation of citation metrics within specific disciplinary contexts. The bibliometric approach is well-suited to the nature of our investigation. It provides a systematic and quantitative way to measure scientific knowledge (Plonsky, 2023), making it useful for exploring the underlying forces that shape the visibility and impact of research outputs.

Guided by this research synthesis method, the case study by Zhang and Aryadoust (2025) examined what factors contribute to the high citation rate of articles in Applied Linguistics. Specifically, in this study, we collected, analyzed, and interpreted the bibliographic data of 302 Quartile-1 journal papers that rank in the top 1% by citations in Applied Linguistics literature (2000–2022). We began with a descriptive analysis illustrating the bibliometric features of the target studies over time, and then conducted a multiple linear regression analysis to assess the relationship between 11 explanatory factors and citation outcomes. Contrary to the notion that citations are a straightforward reflection of scientific merit, we found that citation counts were significantly predicted by 11 extrinsic factors that were independent of research quality: the number of authors, subfield, methodology, title length, CiteScore, accessibility, and scholar *h*-index. Our findings, including both citation drivers and their implications, contribute to a better understanding of how scholarly impact is shaped and assessed within the context of Applied Linguistics.

Beyond its empirical implications for citation-based research evaluation, the study also presents a transparent and systematic methodological process. This chapter discusses it in greater detail in the following sections, demonstrating how bibliometric reviews can be conducted with methodological quality while bridging analytical outputs to theory building and practice improvement. Following the established conventions of research synthesis and bibliometric reviews, the methodological and ethical issues are presented in a step-by-step manner, along with the courses of action to address them. Alternative approaches and their potential implications for the bibliometric research are also discussed. The chapter concludes by highlighting key takeaways, guiding questions for future inquiry, and recommended readings to support deeper engagement with bibliometric reviews.

2. Methodological and ethical issues

At the outset of the study (i.e., Zhang & Aryadoust, 2025), one of the challenges that we faced was the absence of standardized methodological protocols for conducting bibliometric reviews in Applied Linguistics. While disciplines such as business (Donthu et al., 2021) and management (Zupic & Čater, 2015) have developed field-specific guidelines through iterative methodological refinement, Applied Linguistics has only witnessed a recent surge in bibliometric reviews (Plonsky, 2023), leaving much room for methodological advancement. Though useful how-to-do guides on research synthesis in the field are readily available (e.g., Chong, 2025; Norris & Ortega, 2006; Plonsky & Oswald, 2015), they give limited attention to the features of bibliometric methodology and do not offer a framework for predicting citation counts in the fields.

Prior to conducting a bibliometric review that aims to identify the drivers of citation counts, it is crucial to establish a framework with explicit procedures, analytical techniques, and decision-making criteria. This would enhance methodological consistency, transparency, and reproducibility, all of which are cornerstones for producing reliable and interpretable findings (Öztürk et al., 2024). Given that methodological choices are often layered with ethical dimensions, such a framework also serves as a safeguard for research integrity. Therefore, our primary consideration was the development of an applicable methodological framework comprising clearly defined stages of research as outlined below.

2.1 Study design

To design a bibliometric reviewit, is important to address two key points: the specific research problems being investigated and the defined scope of the study. These two steps are discussed next.

2.1.1 *What research problems does the analyst aim to address?*

The initial stage of a bibliometric review involves establishing the study's aim(s). There are two general aims in bibliometric reviews: the evaluative and the relational purposes (Sooryamoorthy, 2021). The evaluative aim is mainly to compare the performance of scientific actors (e.g., journals, authors, articles, institutions, and countries) in terms of their productivity and impact. The relational aim, on the other hand, can be achieved through science mapping tools, which help examine the relationships within research, including the structure and dynamics of research fields, and the patterns of co-authorship. Aim ambiguity arises when researchers fail to make a clear distinction between these two concerns and their respective methodological solutions (Öztürk et al., 2024). For example, in studies seeking to reveal the intellectual structure of a scientific discipline, reliance

solely on evaluative indicators such as publication numbers and citation counts would be essentially limited. Instead, science mapping tools (e.g., co-citation analysis, bibliographic coupling, co-word analysis, and co-authorship analysis) are methodologically relevant in this context (e.g., Aryadoust, 2020, 2023). On the other hand, to address questions like "Which journal has the greatest influence?", "Who are the most prolific authors?", and "What are the highly cited publications?" (Aryadoust & Zhang, 2025, p. 4), performance analysis is a better choice, as it provides deeper insights into the productivity and impact of scientific contributors within the field being studied. As such, the "methodological fit" between research aim(s) and bibliometric approaches should be deliberately defined and clearly articulated.

2.1.2 *What is the scope of the bibliometric review?*

Defining the scope is another methodological dilemma because every decision — how broad or narrow — can have trade-offs that affect the study's generalizability and interpretability. As proposed by Donthu et al. (2021), the scope of a study should be broad enough to justify the use of bibliometric approaches, as these are inherently designed to process and interpret large-scale datasets. However, defining clear boundaries, such as the analytical unit and the time window, ensures that the study is focused and manageable. The basic unit of analysis for bibliometric reviews is documents (Sooryamoorthy, 2021), which take various forms, such as journal articles, book chapters, reviews, conference proceedings, etc. It can also proceed with different levels such as individuals, research teams, institutions, countries or regions. This multiplicity in selecting the subject of inquiry requires meticulous consideration to ensure alignment with the study's aim(s), and each choice carries implications for the depth and breadth of the analysis. Regarding the time coverage of a review, researchers must decide whether to analyze data as a static "snapshot" of a field at a specific time point or to segment the timeline into slices to trace its evolutionary patterns and dynamic shifts (Aria & Cuccurullo, 2017). This temporal flexibility empowers bibliometric reviews to capture both the current state of a research field and its developmental trajectory over time. Importantly, by applying specialized time series clustering techniques, the analyst can trace the emergence, development, and decline of research clusters over time (see Aryadoust et al., 2020; Zakaria & Aryadoust, 2024).

2.2 Data collection

Data collection should address several critical considerations, including the choice of database for the literature search, strategies for ensuring a transparent and reproducible review process, and methods for managing misspellings, dupli-

cates, and incomplete records in the dataset. In the meantime, they are ethical imperatives in nature, which help prevent systematic biases and reduce risks of misinterpretation (also see Chapter 5, Section 2.1). Each of these considerations will be discussed next.

2.2.1 *What database will be used for literature search?*

The successful completion of a bibliometric review largely depends on the availability of bibliographic data that suits its aim(s) and scope. This underscores the critical role of data sources. Numerous available databases can provide the metadata on scientific documents for bibliometric reviews, e.g., Scopus, the Web of Science, Google Scholar, PubMed, Microsoft Academic, Dimensions, and Crossref. Researchers must navigate these options with caution, as limited database coverage can lead to an inaccurate or biased representation of the studied field. Comparative analyses of database performance across disciplines have been extensively documented (Visser et al., 2021). However, it remains a demanding task for applied linguists to decide which database is best suited for their field or study. While general frameworks for evaluating bibliographical databases provide valuable insights, which emphasize criteria such as availability, comprehensiveness, accuracy, and relevance (Pranckutė, 2021; Sooryamoorthy, 2021), our study faced the challenge of adapting these principles to the unique demands of our research context (see Zhang & Aryadoust, 2025).

2.2.2 *How can transparency and reproducibility in the review process be ensured?*

Bibliometric reviews may lack clear documentation of their data collection process, making it difficult for others to replicate their findings (Öztürk et al., 2024). Research transparency and reproducibility are not just methodological concerns; they are core components of responsible conduct of research. It is thus pivotal to ensure all steps are articulated in a way that readers can follow. As Linnenluecke et al. (2020) stated, bibliometric reviews demand rigor akin to other types of research synthesis: researchers cannot arbitrarily include publications but demonstrate a structured and defensible process for identifying and selecting papers. While certain methodological flexibility exists in literature search strategies — there is no universal "right" or "wrong" approach — every decision must be transparent and scientifically supported for better research reproducibility. We believe this rule applies to all data-gathering procedures, where researchers should explicitly detail their methodological choices, including the rationale for database selection, search terms, inclusion and exclusion filters, coding schemes, as well as any encountered limitations.

2.2.3 *How should misspellings, duplicate records, and incomplete data be handled?*

Since bibliographical databases are not free from errors, data cleaning is another critical issue that must be dealt with. Inaccuracies in the publication metadata frequently occur regarding misspelled names of authors, journals, and institutions, incorrect information on publication years, volumes, issues, and pages, as well as errors in citation counts and cited references (Pranckutė, 2021). Incomplete publication records present great barriers, too, especially when the absence of complete author lists in collaborative works and missing institutional affiliations fundamentally undermine studies focusing on research collaboration patterns (Sooryamoorthy, 2021). Databases may also incorporate working papers alongside published versions without systematic duplication or fail to filter out low-credibility sources such as predatory journals (Lim et al., 2024). To mitigate these challenges, therefore, one must consider the inherent limitations of any selected database. This awareness also necessitates implementing rigorous data preprocessing prior to analytical workflows, which often involves tasks such as correcting errors, handling missing information, and removing duplicates.

2.3 Data analysis

The next step in bibliometric reviews involves data analysis, which should consider key methodological choices, such as the selection of appropriate statistical techniques and the approach to graphically presenting the results.

2.3.1 *Which statistical techniques will be selected?*

Bibliometric reviews generate quantitative insights into bibliographic data using a variety of statistical techniques (Aryadoust, 2023). One's choice of analytical approaches, whether descriptive or inferential, should be strategically constrained, employing exclusively those that demonstrate substantive relevance to addressing the defined research questions. The most straightforward way to initiate a bibliometric review could be the descriptive analysis, which illustrates the contributions of scientific actors to a particular field. Much like how empirical research often begins with a summary of descriptive statistics before progressing to deeper analysis, an overview of scientific literature with its components analyzed and reported in a descriptive way can be commonly observed in bibliometric reviews (Donthu et al., 2021). Statistical tools in this type of reviews can include complex methods, such as multivariate or path analysis (see Zhang & Aryadoust, 2026, for an example), which examines multiple variables simultaneously to uncover patterns or relationships within bibliographic data (Soorya-

moorthy, 2021). In such cases, the emphasis shifts to selecting the most relevant variables. This is particularly crucial in predictive research aimed at identifying the drivers of citations. For instance, in regression or path analysis, the careful inclusion of predictor variables — avoiding the omission of key factors and the inclusion of irrelevant ones — is essential for ensuring both the validity and the interpretability of the results (Zhang & Aryadoust, 2025, 2026).

2.3.2 *How will the results be presented graphically?*

Analytical outputs can be presented in intuitive graphical forms, improving the interpretability of complex data and effectively communicating messages to diverse audiences (Passas, 2024). This dual advantage makes visual analytics a common choice in bibliometric reviews. When planning the visual presentation of research findings, one must carefully select the visualization method and the software tool that align with methodological requirements and data characteristics. For basic charts like annual publication trends or citation accumulation curves, general statistical software like R, Jamovi, JASP, SPSS, and STATA (or simply Excel) remain practical options. Their customizable axes, labels, and formatting capabilities allow researchers to present data in an accessible and visually engaging manner. However, specialized bibliometric visualization tools provide superior capabilities for more complex analytical tasks. Popular options include Biblioshiny, SciMAT, CiteSpace, CitNetExplorer, and VOSviewer (for a comprehensive list, see Moral-Muñoz et al., 2020; see also Zakaria & Aryadoust, 2024, for relevant tools used in Applied Linguistics). These tools are particularly advantageous for mapping studies due to their advanced network processing algorithms, built-in normalization techniques, and interactive features enabling dynamic data exploration (Aria & Cuccurullo, 2017).

2.4 Results interpretation

Finally, results interpretation involves connecting the quantitative findings to broader insights and considering their implications for both theory and practice.

2.4.1 *How can quantitative results be linked to meaningful insights?*

The aims of bibliometric reviews directly shape how researchers interpret their results. This echoes Zupic and Čater's work (2015), in which they argued that bibliometric reviews generally fall into three categories based on their primary aims, focusing on "structure, dynamics, or a narrow research question" (p. 448). Our study reported in Zhang & Aryadoust (2025) exemplifies the third category through its inquiry into extrinsic factors that contribute to highly cited Applied Linguistics articles. It is thus essential to critically assess whether the analytical

and visual outputs are in line with such a focus. In other words, the analyst should fully explain how well their findings support the intended research aim(s) and how well they address the research question(s) of the study. In addition, effective interpretation relies on the analyst's ability to situate quantitative findings within disciplinary nuances (Öztürk et al., 2024). To increase the analytical depth, the findings necessitate contextualized interpretation through the lens of Applied Linguistics with domain-specific knowledge. This also means that the analyst should emphasize not just what the data shows, but how it advances the understanding of the field's status quo and unresolved challenges.

2.4.2 *How can the bibliometric review contribute to both theoretical understanding and practical applications?*

Bibliometric reviews have long been critiqued for failing to integrate the findings with theoretical advancement and practical implications (Hulland, 2024). Many previous studies tend to be overly descriptive without offering adequate interpretation or compelling insights to guide future inquiry. While the descriptive portrayal of a scholarly landscape holds value, its true potential emerges only when combined with synthetic wisdom that interrogates gaps and inspires new directions. Accordingly, Mukherjee et al. (2022) proposed ten ways for bibliometric reviews to contribute effectively, giving us much food for thought. As they suggested, the theoretical merits of bibliometric reviews can be achieved if the study (1) discovers knowledge clusters objectively, (2) unpacks nomological networks, (3) analyzes social dynamics in knowledge production and dissemination, (4) traces the field's developmental trends, or (5) highlights unresolved gaps to inform future exploration. Meanwhile, the study can fulfill practical values by (1) evaluating research productivity and impact objectively, (2) establishing boundaries for coverage claims, (3) discovering social dominance or hidden biases for improvement efforts, (4) observing irregularities for further investigation, or (5) assessing performance to support fair and informed decision-making. Mindful results interpretation as drawn from these approaches is thus crucial for our bibliometric review to obtain both scholarly and real-world impact.

3. Course of action

To address the lack of standardized bibliometric protocols for predictive modeling in Applied Linguistics, our study developed a methodological framework (see Figure 1) by integrating interdisciplinary bibliometric guidelines and adapting established research synthesis practices. Specifically, we drew inspiration from the collection of work on research synthesis (Norris & Ortega, 2006; Plonsky &

Oswald, 2015), which provided the systematic procedures and key considerations. Also guided by the general principles in bibliometrics (Donthu et al., 2021), we made field-specific adjustments in data collection to ensure a good representation of prominent Applied Linguistics publications over time and interpreted the results with attention to disciplinary trends and nuances. We believe this integration not only enhances methodological transparency and reproducibility but may also provide a useful framework for future bibliometric reviews to balance quantitative rigor with contextual relevance. To operationalize the proposed framework, the following section elaborates its core components and targeted solutions to address the methodological challenges mentioned earlier. The rationale for each course of action is also provided.

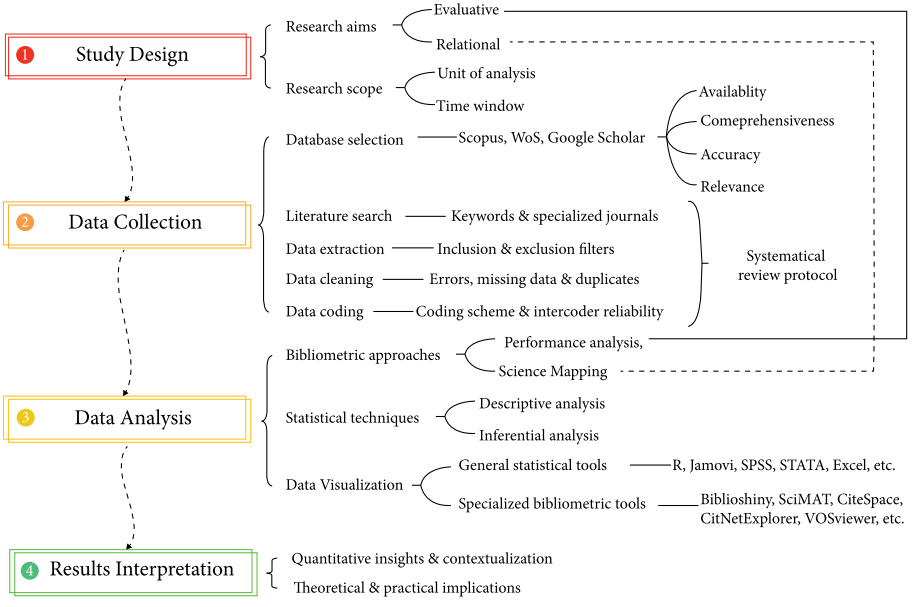


Figure 1. The methodological framework for our bibliometric review

3.1 Study design

3.1.1 *Aim setting*

To carry out our study (Zhang & Aryadoust, 2025), we began by defining a clear evaluative purpose. While citation counts are widely recognized as a proxy for scholarly influence, often shaping the reputation of publications and their authors within a field, significant gaps remain in understanding the complexities that drive citation dynamics (Aksnes et al., 2019). To verify the hypothesis that cita-

tion frequency is a function of various influencing factors besides scientific merit (Garfield, 1972), particularly in the field of Applied Linguistics, we examined the extrinsic factors that may not be necessarily related to research quality but could be significantly associated with article citations. In doing so, our ultimate goal was to better understand how scholarly influence is shaped in Applied Linguistics, and to highlight the importance of evaluating research impact with greater objectivity and rigor.

We further refined our aim into two research questions: (1) What are the bibliometric features of highly cited papers in Applied Linguistics published in the past 22 years? and (2) What extrinsic factors influence the citation counts of these highly cited papers in Applied Linguistics? Aligning this research aim with concrete research questions helped us establish a focused roadmap for what to explore next.

3.1.2 *Scope definition*

The scope of the case study focuses on the top 1% highly cited articles published between 2000 and 2022 from 55 Quartile-1 (Q1) Applied Linguistics-related journals. Specifically, journal articles were selected as our unit of analysis, as they are key sources for disseminating influential research. Limiting the selection to Q1 journals, which are widely recognized for high-quality work, helps control for the effect of research quality and allows the analyst to stay focused on extrinsic factors. The chosen timeframe in this study captures over two decades of research, allowing for an in-depth examination of both historical developments and recent advancements in Applied Linguistics.

Another challenge when defining our research scope was determining the threshold for highly cited papers. By consulting prior research and established citation benchmarks, we took the "top 1%" as the vanguard for impactful scholarship. These choices were based on clear and transpararant criteria, so that we could work with a manageable dataset, and more importantly, capture the most representative research that has greatly shaped and advanced the Applied Linguistics field.

3.2 Data collection

3.2.1 *Database selection*

To ensure a comprehensive source of data, we compared major bibliographic databases and selected Scopus for our study. This decision was informed by prior research highlighting Scopus's broader coverage, particularly in the social sciences and Applied Linguistics domains (Aksnes et al., 2019; In'nami & Koizumi,

2010). To empirically verify this, we conducted parallel searches using the query "applied AND linguistics" across both Scopus and the Web of Science, limiting results to the 2000–2022 period. The results were striking: Scopus yielded 219,730 records, whereas the Web of Science retrieved only 19,008. This substantial disparity underscored Scopus's superior inclusivity and representation within the target domain. Moreover, each article indexed in Scopus comes with various types of meta-data, covering citation information, bibliographical information, abstract and keywords, etc. The database's advanced search functionalities and export options made it well-suited for our investigation (Pranckutė, 2021). As such, choosing Scopus is consistent with the study's methodological commitment to the breadth, relevance, and availability of data in exploring Applied Linguistics scholarship.

3.2.2 *Research transparency and reproducibility*

A transparent and reproducible review was performed in this study by following a clearly structured and documented methodological protocol. Accordingly, a full account of our methodological approach was offered, including scope definition, journal and database selection, inclusion and exclusion criteria, data extraction, coding procedures, and data analysis.

A notable feature of the study is the application of the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 statement, particularly in the reporting of the data collection process. We explicitly stated our search query and filtering criteria (e.g., document type, language, publication years), and further provided a PRISMA diagram to visualize each step of the inclusion process. Data coding procedures were also meticulously developed and executed, with all dependent and independent variables precisely defined, quantified, and grounded in the existing literature. For example, the choice to normalize citation counts using citations per year as the dependent variable shows our awareness of age-related citation bias. While we acknowledged the subjectivity in categorizing an article's subfield or methodological orientation, this was mitigated through the application of a well-calibrated procedure, as detailed in the supplemental materials. Importantly, we applied a rigorous intercoder reliability process to ensure consistency and accuracy in data coding and resolved discrepancies through discussion and refinement of the coding manual. In all, these methodological choices, explicitly laid out in both the main text and supplementary appendices, underscore our endeavor for research rigor and transparency in the study.

3.2.3 *Data accuracy*

After extracting bibliographic records from Scopus, we identified inherent limitations in database metadata, particularly in author affiliations, open access status, and funding information, which are sometimes inconsistently reported or entirely missing. To handle these issues, we downloaded each of the 302 target articles and thoroughly read their content to cross-validate the data. This manual process helped identify and correct typographical errors in author names and institutional affiliations, thus enabling accurate coding of variables such as internationality and geographical origin.

Special attention was given to incompleteness or misclassification in open access and funding information, as Scopus does not always correctly index these details. Duplicate entries were also removed based on title screening to ensure that each article in the dataset was unique and relevant to the study scope. Through this rigorous validation process, we mitigated potential biases arising from metadata inaccuracies and enhanced the reliability of our dataset.

3.3 Data analysis

3.3.1 *Statistical techniques*

To address the two research questions, we employed a combination of descriptive and inferential statistical techniques. To address Research Question 1, which aimed to identify the bibliometric features of highly cited articles within the field, a detailed descriptive analysis was conducted. It involved examining the chronological distributions of citations and the statistical properties of the studied variables. This initial stage was essential not only for an overview of citation patterns and scholarly practices in Applied Linguistics, but also for an empirical foundation upon which more complex inferential analyses could be meaningfully built.

To answer Research Question 2, we conducted a multiple linear regression analysis. This statistical method is suitable for simultaneously assessing the impact of multiple predictor variables on a single outcome, making it appropriate for our investigation to determine the extent to which extrinsic factors influence citation counts. We further applied a logarithmic transformation to correct for excessive skewness in the dependent variable and tested the assumptions of linear regression, including linearity, homoscedasticity, normality, and multicollinearity. Moreover, the use of standardized and sheaf coefficients allowed us to compare the relative predictive strength of individual and grouped variables. In sum, these statistical approaches provided reliable answers to the study's guiding research questions and facilitated an empirically grounded investigation into the social and visibility dynamics shaping citation practices within the studied domain.

3.3.2 *Graphical presentations*

We utilized a variety of visualization methods to present and interpret the patterns found in highly cited Applied Linguistics articles. First, temporal trends were analyzed through two complementary graphics: a scatter plot mapping citation counts of each studied article against publication years to highlight a right-skewed distribution favoring earlier publications, and a histogram illustrating annual aggregated citation totals to identify peaks and declines. Next, frequency distributions for continuous variables, namely, citations per year, author counts, and reference counts were depicted, which highlighted right-skewed data and justified methodological choices like logarithmic transformations. Thirds, categorical variables, including accessibility, internationality, funding, geographical origin, subfield, and methodology, were presented using bar and pie charts, enabling clear comparisons of proportions. These visuals not only characterize the overall citation landscape in Applied Linguistics over the past two decades, but also establish a synergistic relationship between computational analytics and interpretative clarity.

3.4 Results interpretation

3.4.1 *From numbers to insights*

To bridge quantitative results with meaningful interpretation, we synthesized our findings by situating them within the disciplinary context of Applied Linguistics. The study showed that normalized citations were significantly predicted by the number of authors, subfield, methodology, title length, CiteScore, accessibility, and scholar *h*-index. Rather than merely reporting these associations, we contextualized them within broader disciplinary discourses. For instance, we suggested that the higher citation rates observed in the "language contact" subfield may be attributed to the current prominence of the translanguaging research theme in multilingual education (Lin & Lei, 2020). Similarly, we explained why research synthesis articles attracted more citations, given the growing methodological sophistication and the increasing value placed on secondary research within Applied Linguistics (Chong & Plonsky, 2023). In discussing variables like scholar *h*-index, we went beyond numerical patterns to engage with underlying scholarly mechanisms such as the Matthew Effect and the social constructivist nature of citation behavior. We also considered publishing practices unique to the field, such as the prevalence of solo or small-team authorship and the limited uptake of open access despite its potential to increase research visibility. In this regard, the study moved beyond surface-level statistical associations and critically analyzed how broader epistemological and social dynamics influence citation practices in Applied Linguistics.

3.4.2 *Theoretical and practical implications*

From a theoretical standpoint, the present review contributes to both citation theories and scholarship by highlighting the cognitive biases and social dominance in knowledge dissemination (Mukherjee et al., 2022), as reflected by the extrinsic citation drivers and their underlying social mechanisms. In particular, it supports the social constructivist approaches to citation analysis, which regard citations not merely as indicators of academic merit, but as socially constructed acts shaped by strategic and rhetorical scholarly practices (Aksnes et al., 2019).

In addition, identifying significant gaps, such as the underrepresentation of open-access articles and methodological studies in Applied Linguistics, provides a grounded basis for rethinking how academic value is assigned and how future research agendas can be better situated to address visibility biases. In terms of practical implications, the study contributes to ongoing discussions on evaluative bibliometrics, particularly with regard to fairness and methodological transparency, and offers guidance for more informed and equitable research evaluation. We acknowledge the limitations of citation-based metrics, and advocate for a more holistic evaluation approach in Applied Linguistics incorporating both quantitative and qualitative measures. This awareness encourages scholars and policymakers to interpret citation metrics more cautiously and contextually.

4. Alternative courses of action

Several alternative methodological and analytical pathways are available for this bibliometric review, which are presented stage-by-stage below. Exploring these alternatives would shed light on the methodological flexibility in bibliometric reviews by illustrating how varying approaches, such as different units of analysis, data sources, or analytical techniques, might influence the patterns observed, the trends highlighted, and the conclusions drawn. Embracing these possibilities also answers the call to drive ongoing methodological innovations in Applied Linguistics (Plonsky, 2023), and invites a broader dialogue about the best research practices in bibliometric reviews (Öztürk et al., 2024).

4.1 Study design

The present study focused mainly on extrinsic factors — those independent of the scientific quality of an article — to account for variations in citation counts. An alternative path would involve exploring the intrinsic dimensions that fundamentally characterize research excellence, as they are commonly recognized as the

main drivers of citation impact (Kousha & Thelwall, 2024). Research quality is a multidimensional construct, encompassing key attributes such as methodological rigor and plausibility, originality and novelty, as well as scientific and societal impact (Aksnes et al., 2019). Incorporating these intrinsic factors could yield a more comprehensive explanatory model, for researchers to disentangle the extent to which citations reflect genuine intellectual contributions. This endeavor, while supporting our original findings, would also provide empirical evidence for the arguments surrounding the normative versus the social constructivist theory of citations (Bornmann & Leydesdorff, 2015). Nonetheless, operationalizing intrinsic quality would introduce substantial methodological complexity, as the assessment of research quality is highly subjective, and influenced by the heterogeneity of quality standards across fields and disciplines (Thelwall et al., 2023).

4.2 Data collection

This study utilized Scopus as the sole source of bibliometric data, a choice that has been well justified (see Section 3.2). One step further, a more comprehensive data collection strategy that integrates additional sources could strengthen the quality of our dataset. Combining multiple none-overlapping databases would allow for a more inclusive representation of scholarly outputs, as different platforms often index distinct sets of publications. However, the benefits of expanded coverage come with trade-offs. Specifically, merging datasets from multiple sources would introduce challenges such as duplicate records, inconsistent formats, and differences in citation indexing practices. In this regard, a useful procedural guideline was provided by Lim et al. (2024) for integrating bibliometric data from multiple databases, which highlighted the role of specific bibliometric software tools (e.g., Bibliometrix and Biblioshiny) alongside manual efforts in facilitating this process. Meticulous data cleaning processes are thus required to ensure the validity and reliability of the final dataset.

4.3 Data analysis

We employed multiple regression analysis to identify predictors of citation counts. Alternative analytical strategies, particularly those grounded in machine learning, share similar patterns with regression but outperform with large-scale content-based and bibliometric data (Fu & Aliferis, 2010). Algorithms like decision trees, random forests, support vector machines, and K-nearest neighbors are capable of modeling non-linear relationships and capturing complex interaction effects that traditional regression models might overlook (for an in-depth review, see Aiza et al., 2023). Recent advances in deep learning have further propelled

this line of research, with neural network-based approaches gaining traction due to their ability to optimize numerous parameters, exhibit high robustness, and extract high-quality features (Li et al., 2022). Since machine learning algorithms typically require large, high-quality datasets to perform effectively, we need to replan our study scope and data strategy accordingly. In addition, the increased complexity of these models often comes at the cost of interpretability, posing potential challenges when we attempt to translate the synthetic findings into meaningful insights.

4.4 Results interpretation

Given that quantitative outputs may not fully capture the complexities underlying citation behaviors, the interpretation of citation patterns can be enriched through qualitative inquiry, such as interviews or surveys with the authors of highly cited articles. "An insider's perspective" on the citation process could be adopted, illuminating those deliberate strategies for greater visibility, as well as authors' reflections on what they believe contributed to their article's impact (Tahamtan & Bornmann, 2019). Such a qualitative approach, with valuable contextual and explanatory depth, would give more credibility to the inferences we had drawn from bibliographic data and statistical modeling. While issues such as recruitment bias, memory limitations, and subjectivity in interpreting narratives may exist (Aksnes et al., 2023), triangulation could be a critical strategy to enhance the validity and reliability of findings in such contexts.

5. Conclusion

This chapter has charted a dual methodology to support bibliometric reviews in Applied Linguistics. First, by explaining each step of our methodological process, we demonstrated how bibliometric rigor can be systematically operationalized to yield reliable and replicable scientific claims. Notably, we argued that research transparency is both a methodological and ethical requirement, guided by well-designed methodological frameworks and explicitly articulated rationales for each choice made. Second, we underscored the importance of connecting bibliometric outputs to broader theoretical and practical implications, a step often overlooked in previous studies. Our case study illustrated that the true value of bibliometric reviews emerges only when quantitative data is meaningfully interpreted through a contextual lens, illuminating both research gaps and future directions within the studied domain. As a whole, by reflecting on our process and outcomes, this chapter aims to provide a roadmap for researchers seeking

to conduct bibliometric reviews that are not only methodologically sound but also theoretically and practically meaningful. This echoes Mukherjee et al. (2022) that all forms of scholarly research, including research synthesis, play a vital role in advancing the body of knowledge, as long as they are grounded in rigorous design, transparent reporting, and theoretical and practical relevance.

6. Questions for discussion

1. How does the bibliometric approach in this case study enhance the understanding of research visibility and impact?
2. What methodological decisions in the case study contribute to its rigor and transparency?
3. Should bibliometric reviews be field-specific?
4. What key components of our methodological framework help ensure that bibliometric reviews are methodologically robust and practically useful?

AI use declaration statement

The authors confirm that no AI tools were used in the preparation, writing, formatting, and editing of this work.

Further reading

doi Aryadoust, V. & Zhang, S. (2025). How to conduct bibliometric and scientometric research in language education? In H. Nesi (Eds.), *International encyclopedia of language and linguistics*. Elsevier.

This article offers a practical guide for conducting bibliometric and scientometric research in the field of language education. Besides introducing key bibliometric methods and tools, it highlights the role of disciplinary contexts and research specificity, making it a useful resource for those seeking to evaluate research performance and map science in their respective fields.

doi Meihami, H., & Esfandiari, R. (Ed.). (2024). *A scientometrics research perspective in Applied Linguistics*. Springer.

This book thoroughly showcases how scientometric methods can be used to analyze and understand research practices and trends in Applied Linguistics. Featuring contributions from a diverse array of international scholars, it represents the state of the art in adopting scientometrics to explore uncharted territories in the field, and offers valuable insights for both early-career and established researchers for future scientometric explorations.

Plonsky, L. (2023). Introducing bibliometrics in Applied Linguistics. *Studies in Second Language Learning and Teaching*, *13*(4), 721–726.

This editorial introduces a special issue dedicated to the use of bibliometrics within the field of Applied Linguistics. It emphasizes the importance of reflecting on the discipline's publication conventions and research practices through systematic and empirical methods, and situates bibliometrics within the broader context of ongoing methodological reform in Applied Linguistics.

Zhang, S., & Aryadoust, V. (2026). Extrinsic drivers of scientific impact in second language assessment: A bibliometric path analysis. *Language Testing*. Advance online publication.

This paper serves as another concrete example of how to conduct bibliometric reviews with methodological rigor and analytical depth. Situated in the context of language assessment, it goes beyond basic descriptive statistics and uses path analysis to investigate the complex interplay of extrinsic factors in shaping the field's scientific impact.

References

Aiza, W. S. N., Shuib, L., Idris, N., & Normadhi, N. B. A. (2024). Features, techniques and evaluation in predicting articles' citations: A review from years 2010–2023. *Scientometrics*, *129*(1), 1–29.

Aksnes, D. W., Langfeldt, L., & Wouters, P. (2019). Citations, citation indicators, and research quality: An overview of basic concepts and theories. *SAGE Open*, *9*(1), 215824401982957.

Aksnes, D. W., Piro, F. N., & Fossum, L. W. (2023). Citation metrics covary with researchers' assessments of the quality of their works. *Quantitative Science Studies*, *4*(1), 105–126.

Aria, M., & Cuccurullo, C. (2017). Bibliometrix: An R-tool for comprehensive science mapping analysis. *Journal of Informetrics*, *11*(4), 959–975.

Aryadoust, V., & Zhang, S. (2025). How to conduct bibliometric and scientometric research in language education? In H. Nesi (Eds.), *International encyclopedia of language and linguistics*. Elsevier.

Aryadoust, V. (2023). Bibliometrics and scientometrics in Applied Linguistics: Epilogue to the special issue. *Studies in Second Language Learning and Teaching*, *13*(4), 925–933.

Aryadoust, V. (2020). A review of comprehension subskills: A scientometrics perspective. *System*, *88*, 102180.

Aryadoust, V., Zakaria, A., Lim, M. H., & Chen, C. (2020). An extensive knowledge mapping review of measurement and validity in language assessment and SLA research. *Frontiers in Psychology*, *11*, 1941.

Bornmann, L., & Leydesdorff, L. (2015). Does quality and content matter for citedness? A comparison with para-textual factors and over time. *Journal of Informetrics*, *9*(3), 419–429.

Chong, S. W. (2025). Synthesis methods and reporting tool (SMART) for research syntheses in Applied Linguistics. *Research Synthesis in Applied Linguistics*, *1*(1), 17–38.

Chong, S. W., & Plonsky, L. (2023). A typology of secondary research in Applied Linguistics. *Applied Linguistics Review*, *15*(4), 1569–1594.

Donthu, N., Kumar, S., Mukherjee, D., Pandey, N., & Lim, W. M. (2021). How to conduct a bibliometric analysis: An overview and guidelines. *Journal of Business Research*, *133*, 285–296.

Fu, L., & Aliferis, C. (2010). Using content-based and bibliometric features for machine learning models to predict citation counts in the biomedical literature. *Scientometrics*, *85*(1), 257–270.

Garfield, E. (1972). Citation analysis as a tool in journal evaluation: Journals can be ranked by frequency and impact of citations for science policy studies. *Science*, *178*(4060), 471–479.

Hulland, J. (2024). Bibliometric reviews — Some guidelines. *Journal of the Academy of Marketing Science*, *52*, 935–938.

In'nami, Y., & Koizumi, R. (2010). Database selection guidelines for meta-analysis in Applied Linguistics. *TESOL Quarterly*, *44*(1), 169–184.

Kousha, K., & Thelwall, M. (2024). Factors associating with or predicting more cited or higher quality journal articles: An Annual Review of Information Science and Technology (ARIST) paper. *Journal of the Association for Information Science and Technology*, *75*(3), 215–244.

Li, X., Tang, X., & Cheng, Q. (2022). Predicting the clinical citation count of biomedical papers using multilayer perceptron neural network. *Journal of Informetrics*, *16*(4), 101333.

Lim, W. M., Kumar, S., & Donthu, N. (2024). How to combine and clean bibliometric data and use bibliometric tools synergistically: Guidelines using metaverse research. *Journal of Business Research*, *182*, 114760.

Lin, Z., & Lei, L. (2020). The research trends of multilingualism in Applied Linguistics and education (2000–2019): A bibliometric analysis. *Sustainability*, *12*(15), 6058.

Linnenluecke, M. K., Marrone, M., & Singh, A. K. (2020). Conducting systematic literature reviews and bibliometric analyses. *Australian Journal of Management*, *45*(2), 175–194.

Moral-Muñoz, J. A., Herrera-Viedma, E., Santisteban-Espejo, A., & Cobo, M. J. (2020). Software tools for conducting bibliometric analysis in science: An up-to-date review. *Profesional de la Información*, *29*(1), 1–20.

Mukherjee, D., Lim, W. M., Kumar, S., & Donthu, N. (2022). Guidelines for advancing theory and practice through bibliometric research. *Journal of Business Research*, *148*, 101–115.

Norris, J., & Ortega, L. (2006). *Synthesizing research on language learning and teaching*. John Benjamins.

Öztürk, O., Kocaman, R., & Kanbach, D. K. (2024). How to design bibliometric research: An overview and a framework proposal. *Review of Managerial Science*, *18*, 3333–3361.

Passas, I. (2024). Bibliometric analysis: The main steps. *Encyclopedia*, *4*(2), 1014–1025.

Plonsky, L. (2023). Introducing bibliometrics in Applied Linguistics. *Studies in Second Language Learning and Teaching*, *13*(4), 721–726.

Plonsky, L., & Oswald, F. L. (2015). Meta-analyzing second language research. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 106–128). Routledge.

Pranckutė, R. (2021). Web of Science (WoS) and Scopus: The titans of bibliographic information in today's academic world. *Publications*, *9*(1), 12.

Sooryamoorthy, R. (2021). *Scientometrics for the humanities and social sciences*. Routledge.

Tahamtan, I., & Bornmann, L. (2019). What do citation counts measure? An updated review of studies on citations in scientific documents published between 2006 and 2018. *Scientometrics*, *121*(3), 1635–1684.

Thelwall, M., Kousha, K., Stuart, E., Makita, M., Abdoli, M., Wilson, P., & Levitt, J. (2023). In which fields are citations indicators of research quality? *Journal of the Association for Information Science and Technology*, *74*(8), 941–953.

Visser, M., Van Eck, N. J., & Waltman, L. (2021). Large-scale comparison of bibliographic data sources: Scopus, Web of Science, Dimensions, Crossref, and Microsoft Academic. *Quantitative Science Studies*, *2*(1), 20–41.

Zakaria, A., & Aryadoust, V. (2024). A scientometric analysis of Applied Linguistics research (1970–2022): Methodology and future directions. *Applied Linguistics Review*, *15*(4), 1595–1619.

Zhang, S., & Aryadoust, V. (2025). Why are some articles highly cited in Applied Linguistics? A bibliometric study. *Studies in Second Language Acquisition*, 1–23.

Zhang, S., & Aryadoust, V. (2026). Extrinsic drivers of scientific impact in second language assessment: A bibliometric path analysis. *Language Testing*.

Zupic, I., & Čater, T. (2015). Bibliometric methods in management and organization. *Organizational Research Methods*, *18*(3), 429–472.

SECTION 4

Meta-analysis

Section overview

What is a meta-analysis?

Zhouhan Jin & Reza Norouzian
University of Macau | Oregon Department of Education

Definition

Meta-analysis can be understood in both a narrow and broad sense (Plonsky & Oswald, 2015). It can be defined narrowly as "a statistical method for calculating the mean and the variance of a collection of effect sizes across studies, usually correlations (*r*) or standardized mean differences (*d*)," or more broadly as "not only these narrower statistical computations, but also the conceptual integration of the literature and the findings that gives the meta-analysis its substantive meaning" (Plonsky & Oswald, 2015, p. 106).

Methodological stages

Conducting a meta-analysis involves several major stages:

1. Defining the research question
   The first critical stage in conducting a meta-analysis is to clearly define the domain of interest and formulate precise research question(s). The question(s) should be specific, measurable, and focused to ensure a meaningful synthesis of findings across studies.
2. Establishing inclusion and exclusion criteria
   This stage involves setting clear criteria for selecting studies to be included in the meta-analysis. Inclusion and exclusion criteria should be based on factors such as study design, population characteristics, outcome measures, and publication type to ensure consistency and reliability in the analysis.
3. Systematic literature search and study selection
   This stage involves conducting a comprehensive search across multiple databases (e.g., ProQuest Dissertations and Theses, Web of Science, PsycINFO) to identify relevant studies. Tools such as PRISMA (Page et al., 2021) can be used to document the selection process.

https://doi.org/10.1075/rmal.17.07jin

4. Data extraction and coding
   Once relevant studies have been selected, the next step is to systematically extract key data and organize it for analyses. This involves designing a coding sheet and double checking by multiple researchers to enhance reliability.
5. Statistical analyses
   The next stage is to conduct statistical analyses to synthesize findings across studies. This involves weighting effect sizes, choosing a meta-analysis model (i.e., fixed-effects or random-effects models), assessing heterogeneity to determine variability among studies, conducting sensitivity analyses and detecting publication bias (e.g., funnel plots, Egger's test).
6. Interpretation and reporting
   The last major stage is to interpret the results, discuss practical implications, limitations, and future research directions.

Software/resources

Several software tools facilitate the execution of a meta-analysis by supporting study selection, statistical analyses and visualization.

1. ASReview
   ASReview (https://asreview.nl/) is an open-source tool to help researchers select relevant studies while reducing screening time and minimizing human error. It is particularly useful in the systematic literature search and study selection stage of meta-analysis.
2. Covidence
   Covidence is another tool for conducting research synthesis, particularly for systematic reviews and meta-analyses. It is widely used for study screening, data extraction, and risk-of-bias assessment.
3. Comprehensive Meta-Analysis (CMA)
   CMA is a computer program for meta-analysis, offering a wide range of statistical models and visualization tools. It is suitable for researchers with varying levels of statistical expertise.
4. R Programming
   R is an open-source programming environment with various meta-analysis packages (e.g., meta, metafor) that allow for extensive customization and advanced statistical modeling.
5. JASP and Jamovi
   JASP and Jamovi are valuable tools for conducting meta-analyses, particularly for researchers who prefer a user-friendly, open-source statistical software. Both provide intuitive interface with built-in modules for statistical analysis, including meta-analytic computations and visualizations.

Examples of meta-analysis publications

Here are five publications that are useful resources for conducting meta-analysis:

Norris, J. M., & Ortega, L. (2007). The future of research synthesis in applied linguistics: Beyond art or science. *TESOL Quarterly*, *41*(4), 805–815.

This publication provides a reflective overview on the evolution of research synthesis, including meta-analysis, in applied linguistics. It outlines the benefits, limitations, and defining characteristics of systematic syntheses, which is valuable for understanding the state of research synthesis methodologies in applied linguistics.

In'nami, Y., & Koizumi, R. (2010). Database selection guidelines for meta-analysis in applied linguistics. *TESOL Quarterly*, *44*(1), 169–184.

This article is a valuable resource for understanding how to effectively select databases for meta-analysis in applied linguistics. It emphasizes the importance of database selection in ensuring comprehensive literature searches.

Plonsky, L., & Oswald, F. L. (2014). How big Is "big"? Interpreting effect sizes in L2 research. *Language Learning*, *64*(4), 878–912.

This article is a valuable reference for understanding meta-analysis, particularly in terms of interpreting effect sizes. It serves as a field-specific guideline that enhances the interpretation of effect sizes, providing researchers with a nuanced understanding in L2 research contexts.

Boers, F., Bryfonski, L., Faez, F., & McKay, T. (2021). A call for cautious interpretation of meta-analytic reviews. *Studies in Second Language Acquisition*, *43*(1), 2–24.

This publication serves as a critical examination of the methodological choices and potential pitfalls in existing meta-analytic reviews.

Norouzian, R., Jin, Z., & Webb, S. (2025). Increasing meta-analytic quality: A multivariate multilevel meta-analysis of note-taking through exposure to L2 input. *The Modern Language Journal*, 1–23.

This meta-analysis employs a multivariate multilevel meta-analysis (3M) approach to increase transparency and accuracy by accounting for dependencies and methodological differences across studies. The reproducible code can be used as a learning tool for novice scholars of meta-analysis.

AI use declaration statement

The authors confirm that no AI tools were used in the preparation, writing, formatting, and editing of this work.

References

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D. et al. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, *372*, n71.

Plonsky, L., Oswald, F. L. (2015). Meta-analyzing second language research. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 106–128). Routledge.

CHAPTER 7

Lessons from a modern meta-analysis
Methodological insights

Zhouhan Jin & Reza Norouzian
University of Macau | Oregon Department of Education

This chapter provides a reflective exploration of conducting a meta-analysis by revisiting Norouzian et al. (2025). Norouzian et al. (2025) looked at the effectiveness of note taking through exposure to L2 input, improving on an earlier study conducted by Jin and Webb (2024). The study retrieved 55 effect sizes from 27 studies to examine how various factors (e.g., time, note-taking treatment type, proficiency) moderate the effect of note taking. It utilized a multivariate multilevel meta-analysis (3M) approach to account for dependencies across studies while incorporating three methodological confounds. This chapter details the methodological dilemmas we anticipated and those we encountered, the alternative approaches we considered, and the resolutions we adopted. We also emphasize our commitment to open science by sharing reproducible code and data. By summarizing the lessons learned from managing these methodological challenges, this chapter aims to provide practical guidance for researchers seeking more transparent, rigorous, and reproducible meta-analyse, and foster discussions on improving research methodologies in the field of second language acquisition.

Keywords: note-taking, multivariate multilevel meta-analysis, research synthesis, methodological challenges, open science

Background

This chapter reflects on a published meta-analysis on the overall impact of note-taking learning through exposure to L2 input (Norouzian et al., 2025). Note-taking is a widely recognized learning strategy used in both first language (L1; Kobayashi, 2005) and second language (L2; Siegel, 2021) educational contexts. It serves two primary functions: the encoding function, which occurs during note-taking and the external storage function, which allows learners to review notes later for reinforcement (Dunkel, 1988). While note-taking has been extensively

https://doi.org/10.1075/rmal.17.07ajin
Available under the CC BY-NC-ND 4.0 license.© 2026 John Benjamins Publishing Company

studied in L1 contexts, its role in L2 learning remains underexplored. This underrepresentation is surprising, given that note-taking could be particularly beneficial in L2 contexts where learners process unfamiliar linguistic input through taking notes and retaining it for future use (Siegel, 2021). Given its potential to enhance comprehension and retention of linguistic input (Jin & Webb, 2024), a deeper investigation into its effectiveness is warranted.

A recent meta-analysis by Jin and Webb (2024) was the first to systematically synthesize the overall effect of note-taking through exposure to L2 input. Their findings indicated a small-to-medium positive overall effect (Hedges' $g = 0.56$, 95% CI [0.24, 0.88]), suggesting that note-taking can enhance L2 learning outcomes. The study also examined various moderating factors that might explain variability in note-taking effectiveness across different studies.

However, like most meta-analyses in Applied Linguistics, their study employed the classic approach to meta-analysis, which assumes that effect sizes are statistically independent. As Norouzian and Bui (2024) have argued, this assumption is often unrealistic in educational research, where study designs frequently involve dependencies such as repeated measures of the same learners or nested data structures. The classic approach also lacks a systematic way to account for methodological heterogeneity, such as variations in study duration and assignment methods (e.g., intact class vs. randomization) across the studies. These methodological differences can potentially influence study outcomes. If left unaccounted for, they can confound interpretations of moderator effects, potentially leading to inaccurate conclusions about the extent to which note-taking is effective. To overcome these challenges, Norouzian et al. (2025) employed a multivariate multilevel meta-analysis (3M) approach to provide a more transparent and robust analysis by accounting for (a) dependencies in the evidence presented by primary studies, (b) methodological differences confounding the effectiveness of interventions, (c) differences in research designs, and (d) enhancing the accessibility of findings by using percentages. In addition to adopting a 3M model, our research synthesis was further guided by the (M)UTOS framework to systematically organize and interpret moderator variables. The conceptual and operational implications of this framework are introduced in detail in the sections that follow.

The meta-analysis synthesized findings from 27 studies, retrieving 55 effect sizes to determine the overall impact of note-taking learning through exposure to L2 input. Results revealed that note-taking was generally beneficial, with at least a 63% likelihood of producing a meaningful learning gain. Among different note-taking strategies, structured approaches such as vocabulary notebooks and framework notes yielded higher learning outcomes than conventional, unstructured note-taking. Furthermore, the effectiveness of note-taking was moderated

by factors such as learners' proficiency level, the type of input mode (reading vs. listening), and the measurement format (recognition vs. recall tests).

These findings provide significant implications for both L2 learners and educators. The study highlights the importance of structured note-taking interventions in language learning and provides empirical support for instructional practices that integrate effective note-taking strategies. Also, this percentage-based interpretation of effect sizes responds to one of the misconceptions discussed in Chapter 1 that research syntheses are useful only for researchers. By translating statistical estimates into percentages, we aim to make the findings more accessible to teachers and other practitioners beyond academia. Additionally, the 3M approach sets a new standard for meta-analytic research in Applied Linguistics, demonstrating the need for methodological rigor in synthesizing data. By shedding light on the conditions under which note-taking is most beneficial, this meta-analysis offers valuable insights for future research and pedagogical applications.

However, during and prior to conducting this synthesis, we encountered a range of methodological and ethical challenges that required careful deliberation and adjustment. The subsequent sections will further explore the methodological and ethical dilemmas encountered in this study, the course of action taken to address them, alternative approaches, and the broader implications for L2 learning and research methodologies.

Methodological and ethical dilemmas

Beyond technical considerations, several of the dilemmas we encountered were ethical in nature. These included decisions about the inclusion of non-peer-reviewed studies (balancing inclusivity with quality control), whether to adjust for intact-class clustering to avoid inflated precision, how to handle sparse delayed posttest data without overstating findings, and how transparently to document analytic decisions through open science practices. Each of these decisions had implications not only for methodological rigor but also for fairness, representativeness, and responsible knowledge production.

Our synthesis began with the recognition that conducting a rigorous meta-analysis in Applied Linguistics would involve navigating multiple methodological and ethical dilemmas. Drawing upon our work written by Norouzian et al. (2025), we systematically anticipated potential issues and documented unexpected challenges that emerged during the synthesis. These dilemmas spanned from developing a conceptual model to finalizing the coding scheme and making inclusion decisions for studies of varying quality. We structure this section by first addressing the issues we anticipated and then those that arose during the process.

Anticipated methodological issues

Challenges in variable definitions and classification

Before commencing the synthesis, we anticipated considerable ambiguity in the definitions of variables used across primary studies. One primary challenge was how to classify moderator variables within a framework that organizes them in a principled way. One such organizing framework is (M)UTOS (Aleo et al., 2020; Norouzian & Bui, 2024). The (M)UTOS framework calls for organizing moderators into methodological (M), unit (U), treatment (T), outcome (O), and setting (S) variables.

But it is researchers that get to use their expert judgment to assign moderators to these categories. In our case, studies often used different terminologies or provided insufficient details about their methodologies. For instance, note-taking approaches ranged from conventional, unstructured note-taking to more structured and strategic note-taking such as framework notes or vocabulary notebooks. However, not all studies explicitly described their instructional procedures or the degree to which students were guided in their note-taking practices. Similarly, learner proficiency is another variable that was inconsistent or absent from the reports. Some studies used standardized tests such as TOEFL, while others provided no explicit information, requiring us to infer proficiency based on contextual cues.

Additionally, we anticipated subjectivity in coding certain study characteristics. For example, we accounted for three nonsubstantive or methodological differences in (a) the length of studies from the pretest to the last posttest (in weeks), (b) the frequency of employing the note-taking treatment, and (c) the overall design (true vs. quasi-experiments) across our studies. In terms of the overall design, it is important to determine whether a study is qualified as a true experiment, where participants or classes are randomly assigned to groups, or as a quasi-experiment, where such randomization is not used. We anticipated that several studies would provide ambiguous methodological details, requiring us to interpret available descriptions cautiously. Since this classification directly impacted the robustness of effect sizes, we understood early on that coder judgment would play a significant role in data extraction and coding reliability.

Low-frequency categories also posed challenges (see Norouzian, 2025). After examining the frequency of each category occurring across the studies, we found that several categories needed to be modified. For example, different L1s such as Japanese only appeared in one study, accounting for a total of two effect sizes, while Chinese appeared in two studies, accounting for eight effect sizes. However, it is recommended that a minimum of three studies is included in each moderator subgroup (Li, 2016; Vuogan & Li, 2023).

Ethical considerations — Inclusion of non-peer-reviewed studies

Another dilemma we anticipated involved balancing the inclusion of a comprehensive dataset with maintaining methodological rigor, which aligns with the discussion in Chapter 8 regarding the inclusion of grey literature and research published in less visible venues. Three of the studies (Amini & Sadati, 2023; Nagep, 2022; Ulfani et al., 2023) retrieved through Google and Google Scholar were published in less visible venues, including non-peer-reviewed journals. These sources raised concerns about quality control and the reliability of reported findings. We therefore carefully evaluated whether the inclusion of such studies could be ethically and methodologically justified.

Unanticipated challenges during the synthesis

Moderators inapplicable across all studies

As we progressed through the coding process, we found that some moderators could not be universally applied across the dataset. For example, while we initially aimed to include a moderator for the average duration (in minutes) of note-taking interventions, this variable was not applicable in observational studies where learners self-selected note-taking strategies without an instructional component.

Adjusting for intact-class assignment

During the synthesis, we found that approximately 70% of the selected studies assigned participants using intact classes rather than individual randomization. This clustering introduced potential bias because data gathered from a group of learners sharing the same class, teacher and instructional context tends to exhibit varying degrees of dependency (see Hedges, 2007; Norouzian, 2025). This dependency or overlap reduces the overall information worth of the collected data for generalizability purposes. Therefore, it is critical to adjust studies that assign participants to groups by intact classes to align them with studies that assign individual students to groups.

Unreliable delayed posttest data

During analysis, we found that only 3 out of 27 studies reported delayed posttest results, yielding just four effect sizes. The limited dataset raised concerns about the robustness of any findings drawn from these data. As a result, we faced a decision point: whether to include these limited effects or exclude them to avoid unreliable results.

Developing a conceptual model

Despite the practical ubiquity of note-taking in educational settings, the conceptual framework explaining its role in L2 learning remains underdeveloped in the literature. Early in our synthesis, we encountered significant uncertainty in conceptualizing a theoretical model that would guide our meta-analysis. Initial versions of this framework were overly complex, potentially overwhelming to readers. Indeed, because so few studies have explicitly theorized or empirically examined how note-taking affected the impact of L2 input, we found limited evidence aligned with our emerging conceptual model.

Balancing moderator inclusion

One key dilemma emerged during the synthesis is to balance the inclusion of a comprehensive set of moderators with the need to maintain statistical clarity, analytical rigor, and adherence to word limits. Although the inclusion of more moderators could theoretically yield a more complete picture of the variability in note-taking effects, including too many moderators would risk model overfitting and interpretive repetition. Therefore, we decided to selectively include certain moderators.

These dilemmas, both anticipated and unanticipated, shaped the course of our synthesis and ultimately strengthened its methodological foundation. In the following section, we detail the specific actions taken to address these challenges.

Courses of action

To respond to the methodological and ethical dilemmas outlined above, we adopted a combination of strategic planning, iterative refinement, and collaborative decision-making. Below, we outline the specific steps taken to ensure methodological integrity and analytic transparency throughout the synthesis.

Refining the coding process

To improve the reliability and consistency of our coding framework, we held extensive discussions to clarify ambiguous definitions, and systematically categorized moderators using the (M)UTOS framework. We created a reference table (see Table 1) to document which moderator fell under each (M)UTOS category, enhancing clarity and ensuring alignment across our dataset.

For ambiguous definitions, such as note-taking type, and learner proficiency, we reviewed the primary texts collaboratively, revisited our definitions, and

Table 1. (M)UTOS for note-taking moderators

No.	Moderator	M	U	O	T	S	Notes
1	study_length	M/					
2	time_to_test	M/					
3	no_treat	M/					
4	true_experiment	M/					
5	time			O/			
6	outcome			O/			
7	measure			O/			
8	test_type			O/			
9	study_setting					S/	
10	lang_context					S/	
11	region					S/	
12	input_mode				T/		
13	material_type				T/		
14	note_option				T/		
15	note_instruct				T/		
16	prof		U/				
17	ortho_script		U/				

refined the classification scheme until full consensus was achieved. For example, we agreed to differentiate note-taking types into distinct categories: conventional note-taking, framework notes, and vocabulary notebooks. However, disagreements emerged regarding the classification of note-taking instruction. After in-depth discussion, we decided to explicitly define this category to reduce future ambiguity. We added a clarifying note stating that note-taking instruction includes but is not limited to instruction on how to use abbreviations, organize notes, recognize organizational cues like numeric identifiers, and attend to the teacher's flow of speech when input is aural. As a result, the note-taking instruction category emphasizes the provision of general guidance on note-taking skills rather than a specific type of note-taking (e.g., framework notes or vocabulary notebooks). These refinements significantly enhanced the transparency, consistency, and interpretability of our coding process and helped establish a shared understanding that we relied on through the rest of the coding and analysis process.

In parallel with refining note-taking classifications, we also took deliberate steps to address ambiguity in the reporting of learners' proficiency levels. Given the considerable variability in how proficiency was defined or reported, we examined each study carefully looking for contextual cues. For example, in Siregar (2022), we inferred an "intermediate" proficiency classification based on participants' pretest scores and the nature of the assessment instruments detailed in the appendix. These measures suggested a moderate command of language proficiency, aligning with our definition of intermediate-level learners. Similarly, in Nagep (2022), although no standardized test scores were reported, we considered "intermediate" to be safe given that English was required by the program and English was the medium of instruction. In studies that included standardized test scores, such as Al-Ghazo (2023), we used well-established cutoffs from instruments like the TOEFL to justify our classification decisions.

Additionally, in order to maintain sufficient group sizes for meaningful comparison, we merged "Chinese" and "Japanese" under the broader category of "East Asian". For variables that were difficult to categorize due to low frequency or ambiguity, we created a "miscellaneous" category. All the subjectivity involved in the coding decision were under multiple rounds of discussion and refinement to resolve coding disagreements. Finally, we established a coding scheme that ensured consistency across studies, minimized subjective interpretations, and allowed for more accurate meta-analytic comparisons.

Use of proxy measures for inconsistent moderators

When facing moderators that are not applicable across all studies, such as instructional length, which is irrelevant for conventional note-taking as it does not involve instruction, it would be useful to use moderators that act as proxies for the original moderator, or those that may overlap with it (Norouzian, 2025). Therefore, we decided to use a proxy, which is a time-related variable — the length of studies from the pretest to the last posttest (in weeks) — to maintain comparability across studies. This approach can also help improve the comparability of studies and enhance the validity of the findings (Becker, 2017).

Statistical adjustments for intact-class designs

Ideally, adjusting for the amount of overlap between the participants coming from the same classes requires knowing the intra-class correlation (ICC) values, which are almost never reported in primary studies. To address this, we adopted a medium-low ICC of 0.15, as suggested by Hedges (2007), to reflect a conservative level of similarity among students within the same class. ICC values range from 0

to 1. Values near 1 suggest that participants within the same class exhibited highly similar performance, indicating strong dependency. In contrast, values near 0 imply that participants performed very differently, reflecting near-independence among observations.

For intact-class adjustments to work, two additional elements are typically required: (1) the total number of intact classes recruited for the entire study and (2) the number of participants in the treatment and control groups. When this information was available, we used the 3m R package (Norouzian & Bui, 2024) to make the necessary corrections to effect sizes and their sampling variances.

To illustrate the adjustment process, Norouzian (2025) used a subset of our dataset named dat_note, derived from our meta-analysis (Norouzian et al., 2025). The dataset includes unadjusted Hedges' *g* effect sizes (g), their sampling variances (v_g), the assignment type (assign_type, e.g., individual or intact class), the number of intact classes (n_class), and the number of participants in treatment (Nt) and control (Nc) groups. A step-by-step guide for implementing these adjustments is available at https://rnorouzian.github.io/m/p.html. The adjusted values, which are reported as g2 (adjusted Hedges' g) and v_g2 (adjusted sampling variance), are presented in Table 2.

Table 2. Hedges' g and its sampling variability adjusted for assignment by intact classes

g	v_g	assign_type	n_class	Nt	Nc	g2	v_g2
0.319	**0.0830**	class	4	25	24	*0.3073*	**0.2206**
0.986	**0.0912**	class	4	25	24	*0.9499*	**0.2306**
0.723	**0.0870**	class	4	25	24	*0.6966*	**0.2255**

Note. Hedges' g estimates are italicized to highlight their minimal change.
Sampling variances of Hedges' g estimates are boldfaced to highlight their substantial change.

As shown, the adjustments more than doubled the sampling variances and slightly decreased the effect sizes, indicating how assignment by intact classes can compromise both the precision and magnitude of observed effects if left uncorrected. These adjustments could reduce the risk of overestimating the precision and generalizability of findings. Although ICC values and clustering details are often omitted in primary studies, we advocate for future researchers to report them clearly. Doing so would enable more accurate meta-analytic corrections and help improve the methodological transparency and reproducibility of synthesis studies involving grouped or nested designs.

Addressing sparse data in delayed posttests

Although our initial plan included an investigation of delayed posttest effects, only three studies offered usable data, providing just four effect sizes. After several rounds of discussion, we determined that including these effects would yield an unreliable effect size summary, given the sparse categorical moderator (i.e..time) for the delayed posttest category (i.e. posttest2). Therefore, we decided to exclude these delayed posttest effects from our narrative interpretations to maintain transparency and avoid drawing conclusions from insufficient evidence. However, we still provided results from delayed posttest (posttest2) using figures and tables within the supplementary materials for readers interested in reviewing the full dataset.

Conceptual model development

We refined our conceptual model (see: https://rnorouzian.github.io/m/p.html #fig) through rounds of discussion. Early versions often attempted to account for more variables than our dataset supported. Through several refinements, we narrowed our focus to key constructs, which are input, note-taking processes, and learning outcomes. We also identified two primary pathways to learning: a direct effect of L2 input on learning outcomes (path c), as previously established in Krashen's (1985) Theory of Comprehensible Input, and an indirect effect where note-taking mediates the impact of input (paths a × b). An additional complexity arose in linking this model to our (M)UTOS moderator framework. The (M)UTOS framework classifies moderators into methodological (M) and substantive dimensions. As we clarified in the supplementary materials, all UTOS variables could potentially moderate either the note-taking process itself (path a) or the processing of L2 input (path c). For example, learners' L1-L2 orthographic distance may affect the ease with which learners can understand the input (Zhang & Zhang, 2022), and in turn affect their ability to take notes. Similarly, the different regions where the study was conducted might also influence learners' note-taking perceptions and habits (Siegel & Kusumoto, 2022).

Ultimately, we developed a conceptual framework that focused on how note-taking mediated the relationship between input exposure and learning outcomes. This final model (see Figure 1) was included in our publication and served as a visual aid to represent the theoretical rationale that guided our analysis.

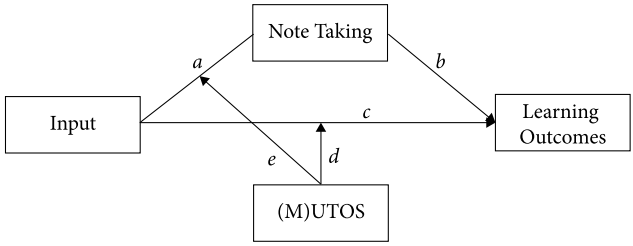


Figure 1. Theoretical framework for note-taking

Approach to moderator inclusion and documentation strategy

To address these concerns, we adopted a selective and principled approach to moderator inclusion. We prioritized moderators using two main criteria:

1. Prevalence in the Dataset: Moderators needed to be present in a sufficient number of studies to allow for meaningful comparisons. Several moderators were excluded due to insufficient or incomplete data. For example, the opportunity to review notes lacked sufficient baseline data to support valid learning gain estimates. Similarly, the moderator on note-taking instruction length could not be reliably analyzed, as it was inherently inapplicable to conventional note-taking treatments and risked introducing coding bias.
2. Redundancy Avoidance: Moderators that overlapped conceptually with others or produced duplicative patterns were excluded. For example, the institutional level variable used in Jin and Webb (2024) as a proxy for proficiency was replaced by a more direct and validated proficiency measure derived from contextual indicators such as TOEFL scores and institutional requirements. In addition, moderators like the number of note-taking sessions overlapped conceptually with methodological variables already included in our model (i.e., frequency of treatment), and were therefore excluded.

Due to the word limit of the manuscript, we chose to provide a detailed account of the decisions behind moderator selection and exclusion in our online supplementary document (https://rnorouzian.github.io/m/p.html). Besides, this comprehensive, publicly accessible online resource includes the complete dataset, R scripts, full model outputs, outlier detection procedures, and additional results not reported in the article. It allows interested readers to examine the full analytic process, verify the findings, and even replicate the meta-analysis independently.

Ethical inclusion of non-peer-reviewed studies

To balance comprehensiveness with quality, we evaluated each non-peer-reviewed study on its methodological transparency, completeness and relevance. After careful examination and discussion, we decided to include these less visible studies. The argument for including such studies was that the authors of these studies are often novice researchers with potentially diverse native language backgrounds. These studies may be rejected by leading L2 journals not due to flawed research design, but because of issues such as lack of novelty or weak academic writing. We argue that neither of these reasons alone is sufficient to exclude such studies from a meta-analysis. For instance, studies with limited novelty may serve as (partial) replications of prior research, thereby increasing the precision in the meta-analytic findings. Furthermore, as long as these studies report the necessary information for meta-analysis such as relevant procedures, moderators' information, and statistical summaries, their writing quality does not interfere with data extraction or synthesis. Other important aspects of the studies, such as their design and data analyses, can still be assessed by meta-analysts on a case-by-case basis. To apply a blanket rule that such studies, by default, should be excluded from a meta-analysis introduces a form of researcher bias to the meta-analysis. This concern parallels the discussion in Chapter 8, which also outlines several practical steps for maintaining methodological quality without disadvantaging studies published in less visible venues. Thus, upon careful examination and in alignment with our inclusion criteria, we decided to include these studies in our meta-analysis. This choice allowed us to expand the empirical scope of our synthesis, enabling a more nuanced understanding of how note-taking influences L2 learning across diverse study designs and participant populations.

Alternative courses of action

Throughout the synthesis, we considered a number of alternative approaches to address the methodological challenges we encountered. While we ultimately selected strategies that best aligned with our goals for rigor, transparency, and clarity, it is important to acknowledge the paths not taken, and to reflect on the potential consequences if we had decided to choose them.

Handling inapplicable moderators in observational studies

In cases where certain moderators, such as the duration of instruction, were inapplicable to observational studies, we considered multiple coding strategies. One option is to leave the moderator blank (coded as NA). However, it would poten-

tially reduce the sample size due to missing data. Another way is to code the instruction length as 0. While coding such values as 0 would allow the inclusion of more studies in moderator analyses, it could misrepresent the construct and inflate the number of studies with 0 values. Because instruction is inherently inapplicable to observational designs, we believed that these studies should not be treated as lying on the same continuum. Ultimately, applying a proxy (i.e., study duration in weeks) as discussed earlier could be a better choice to minimize both missing data and conceptual distortion (Norouzian, 2025).

Use of default assumptions for cluster adjustment

In adjusting for intact-class assignments, we relied on an assumed ICC value of 0.15. An alternative would have been to adopt a more conservative or liberal default, or to exclude cluster-adjusted studies altogether.

Each of these paths presented trade-offs. A lower ICC might have underestimated intra-class similarity, inflating the precision of summary effects (e.g., narrower CIs, and lower p-values). A higher ICC might have been overly conservative, reducing precision excessively. Excluding studies using intact classes would have eliminated the majority of our dataset and introduced selection bias. Our selected ICC offered a balanced and literature-supported compromise.

Inclusion of delayed posttest effects

One alternative we considered was to merge immediate and delayed posttest effects into a single "posttest" category while statistically adjusting for the fact that the delayed posttest studies likely had longer durations. This would have allowed us to retain all available data while accounting for study length as a potential confounding factor. However, although this method could have increased the number of effect sizes, it also risked obscuring differences between immediate and delayed effects, potentially reducing interpretive clarity. We ultimately decided to exclude the limited delayed posttest effects to allow for greater transparency within the results.

More comprehensive conceptual model

When designing the model, the (M)UTOS framework served as a foundation for our conceptualization. We also considered alternative models that could have incorporated a broader range of variables, such as those that inhibit learning or shape learner engagement in more nuanced ways. One such direction could have been to explicitly include variables beyond those that we used in our current

model. For example, learner-related variables such as working memory, attention, or language anxiety might affect the extent to which learners can effectively engage with input and concurrently take notes. Similarly, cognitive overload, particularly in contexts where learners are exposed to dense or unfamiliar academic input, may reduce the effectiveness of note-taking by overwhelming learners' processing capacities. Our current model's exclusion of these variables likely reflects the limited reporting of them in the primary studies. Nevertheless, such variables may play a significant role in moderating the relationship between note-taking and learning.

Additionally, while our analysis focused on the encoding function of note-taking (i.e., the process of note-taking), an alternative analytical pathway could have involved examining the external storage function of note-taking (i.e., review effects). This perspective emphasizes the value of notes that support delayed review and consolidation. Including studies that allow learners to review their notes before posttests could offer insight into the delayed and cumulative benefits of note-taking, especially in longitudinal learning contexts. This would require a modified conceptual model, potentially one that incorporates an additional pathway from note-taking to delayed learning outcomes.

Finally, while our current model treats the cognitive processes underlying note-taking (e.g., encoding) as unobservable, a more fine-grained future approach might attempt to estimate these latent cognitive constructs, provided that relevant proxies (e.g., time-on-task, richness of notes) are available.

Taken together, these alternative courses of action suggest that although the current conceptual model offers valuable insights, other potential models may also be helpful in capturing different dimensions of note-taking and its role in L2 learning.

Retention of all moderators regardless of frequency

Another possible course of action was to retain every coded moderator, even those backed by only one or two studies, and to present all results with equal emphasis. This would have maximized comprehensiveness and potentially highlighted lesser-known influences on note-taking effectiveness.

However, this strategy would be more likely to produce unstable or misleading effect estimates. To avoid drawing conclusions based on insufficient evidence, we chose to prioritize variables supported by broader empirical coverage. That is, we ensured that each moderator subgroup should at least have three studies (Li, 2016; Vuogan & Li, 2023). Nonetheless, we included descriptive summaries of the less frequent moderators in supplementary materials for interested readers.

Embedding all information in the main article

Before initiating our synthesis, we designed a comprehensive coding sheet that aimed to capture a broad range of potential moderators. This extensive coding process reflected our commitment to thoroughly examining the factors that might influence the relationship between note-taking and L2 learning. This comprehensive approach was intended to maximize insights drawn from our synthesis.

However, if we include all of the moderators, we would include more analytical details and diagnostics within the published article. This could either significantly condense other narratives such as background and discussion or sacrifice transparency and clarity in the presentation of results. Embedding such a large amount of information in the main article would overwhelm the reader and reduce the accessibility for replication. The chosen strategy of providing full documentation via a supplementary site preserved both conciseness and openness.

Exclusion of non-peer-reviewed studies

We also considered excluding all non-peer-reviewed studies on the grounds that they may lack rigorous quality control. This would have simplified the synthesis process and reduced the risk of incorporating poorly conducted research.

However, such exclusion could introduce publication and selection bias as discussed in Chapter 1 and so overlook meaningful contributions from specific publication types and less visible sources. Many non-peer-reviewed studies in our dataset provided sufficient methodological detail and data transparency. By evaluating these studies case by case, we aimed to preserve comprehensiveness without sacrificing quality.

Taken together, these alternative paths highlight the importance of careful judgment in synthesis research. Each decision entails trade-offs between inclusivity, validity, clarity, and interpretability. While we believe our chosen course of action best serves the aims of our study, these alternatives offer useful insights and raise important considerations for future meta-analysts.

Conclusion

This chapter has outlined the methodological journey of conducting a multivariate multilevel meta-analysis on note-taking in L2 learning, reflecting both anticipated and unforeseen dilemmas. From coding challenges and conceptual uncertainty to dataset limitations and ethical considerations, the process demanded flexibility, transparency, and sustained critical reflection. Our goal

throughout was to ensure methodological integrity while enhancing the interpretability and practical value of our findings.

Key lessons include the value of early planning and iterative discussion when defining moderators and developing conceptual frameworks. Strategic coding refinement and adjustments for study design (e.g., intact-class assignments) were essential to achieving robust estimates. Furthermore, we found that a balance must be struck between inclusiveness and analytic focus, especially when working within word limits.

Future researchers may benefit from adopting a flexible yet principled approach to conducting meta-analyses, one that remains open to emergent challenges while grounded in theory and best practices. The integration of 3M models with the (M)UTOS framework provided both analytic rigor and conceptual clarity, offering a replicable pathway for future work.

Ultimately, the reproducible code presented in the supplementary materials contributes to open science practices. For those conducting meta-analyses, open science promotes transparency, enables reuse of data and workflows, thereby advancing methodological rigor and innovation in Applied Linguistics. It can also serve as valuable eductional resources for readers and other practitioners. To further embed open science in the meta-analytic workflow, researchers can provide datasets and analysis code in public repositories, and offer detailed supplementary documentation. By adopting these practices, future meta-analyses can become more transparent, reproducible, and ultimately advance knowledge of meta-analysis in the field.

Questions for discussion

1. What strategies can researchers use to decide which moderators to prioritize in a meta-analysis?
2. How can research teams effectively collaborate to resolve coding disagreements and refine conceptual models?
3. How does the use of the 3M approach improve the meta-analytic findings compared to traditional methods?
4. In what ways can supplementary materials or other open science practices be used to enhance the comprehensiveness and transparency of a meta-analysis or research synthesis in general?

Further reading

Norouzian, R. (2025). Advancing meta-analysis in Applied Linguistics: Coding and effect size considerations for improved rigor. *Research Synthesis in Applied Linguistics*, *1*(1), 53–73.

This paper focuses on enhancing the methodological rigor in meta-analyses within the field of Applied Linguistics, particularly through improved coding practices and detailed consideration of effect sizes. It integrated theoretical insights with practical challenges for conducting robust meta-analyses, which are crucial for synthesizing research findings and advancing the field effectively.

Norouzian, R., & Bui, G. (2024). Meta-analysis of second language research with complex research designs. *Studies in Second Language Acquisition*, *46*, 251–276.

This paper introduces why complex research designs in SLA create dependency problems for meta-analysis, and how to handle such dependencies, making it especially valuable for researchers looking to move beyond traditional meta-analytic models toward more robust, transparent approaches.

Norouzian, R., Jin, Z., & Webb, S. (2025). Increasing meta-analytic quality: A multivariate multilevel meta-analysis of note taking through exposure to L2 input. *Modern Language Journal*, *109*(1), 171–193.

This study applies a multivariate multilevel meta-analysis (3M) approach to synthesize research on note-taking in L2 learning, accounting for dependencies and methodological variation across studies. It serves as a detailed, practical example of implementing advanced meta-analytic methods with fully reproducible code and materials, making it a practical example of improving transparency and rigor in meta-analyses.

Siegel, J. (2021). *Developing notetaking skills in a second language: Insights from classroom research*. Routledge.

This book examines how L2 learners develop note-taking skills in real classroom settings, integrating empirical research with practical guidance for instruction. It provides valuable context on the pedagogical side of note-taking, helping researchers and teachers understand how note-taking interventions can be designed and evaluated in authentic learning environments.

Boers, F., Bryfonski, L., Faez, F., & McKay, T. (2020). A call for cautious interpretation of meta-analytic reviews. *Studies in Second Language Acquisition*, *43*(1), 2–24.

This paper critiques common pitfalls in interpreting meta-analytic results in Applied Linguistics, emphasizing issues like heterogeneity, moderator analyses, and publication bias. It advocates for a more nuanced reading of meta-analytic findings, offering practical advice for those who want to conduct a meta-analysis.

AI use declaration statement

During the preparation of this work, the first author used ChatGPT for proofreading, mainly to improve grammar, wording, and overall fluency. As authors, we confirm that after using the tool, we reviewed and edited the content as needed and take full responsibility for the content of the publication, in alignment with COPE and the publisher's AI policy.

References

Al-Ghazo, A. (2023). The impact of note-taking strategy on EFL learners' listening comprehension. *Theory and Practice in Language Studies*, *13*(5), 1136–1147.

Aloe, A. M., Thompson, C. G., & Reed, D. K. (2020). CUTOS: A framework for contextualizing evidence. In B. Albers, A. Shlonsky, & R. Mildon (Eds.), *Implementation science 3.0* (pp. 39–52). Springer.

Amini, S., & Sadati, Z. (2023). The contribution of Cornell note-taking strategy instruction to the listening comprehension of Iranian EFL learners: A case of learners' perception. *Biannual Journal of Education Experiences*, *6*(1), 169–187.

Becker, B. J. (2017). Improving the design and use of meta-analyses of career interventions. In J. P. Sampson, E. Bullock-Yowell, V. C. Dozier, D. S. Osborn, & J. G. Lenz, *Integrating theory, research, and practice in vocational psychology: Current status and future directions* (pp. 95–107). Florida State University.

Dunkel, P. (1988). The content of L1 and L2 students' lecture notes and its relation to test performance. *TESOL Quarterly*, *22*, 259–281.

Hedges, L. V. (2007). Effect sizes in cluster-randomized designs. *Journal of Educational and Behavioral Statistics*, *32*, 341–370.

Jin, Z., & Webb, S. (2024). The effectiveness of note taking through exposure to L2 input: A meta-analysis. *Studies in Second Language Acquisition*, *46*, 404–426.

Kobayashi, K. (2005). What limits the encoding effect of note-taking? A meta-analytic examination. *Contemporary Educational Psychology*, *30*, 242–262.

Krashen, S. D. (1985). *Second language acquisition and second language learning*. Pergamon. (Original work published 1981).

Li, S. (2016). The construct validity of language aptitude: A meta-analysis. *Studies in Second Language Acquisition*, *38*, 801–842.

Nagep, D. S. M. (2022). The impact of five note-taking techniques on academic listening comprehension of EFL student teachers. 57–1 ,(253)22 ,مجلة القراءة والمعرفة.

Norouzian, R. (2025). Advancing meta-analysis in Applied Linguistics: Coding and effect size considerations for improved rigor. *Research Synthesis in Applied Linguistics*, *1*(1), 53–73.

Norouzian, R., & Bui, G. (2024). Meta-analysis of second language research with complex research designs. *Studies in Second Language Acquisition*, *46*, 251–276.

Norouzian, R., Jin, Z., & Webb, S. (2025). Increasing meta-analytic quality: A multivariate multilevel meta-analysis of note taking through exposure to L2 input. *The Modern Language Journal*, *109*(1), 171–193.

Siegel, J. (2021). *Developing notetaking skills in a second language: Insights from classroom research*. Routledge.

Siegel, J., & Kusumoto, Y. (2022). A cross-cultural investigation of L2 notetaking: Student habits and perspectives. *Journal of Multilingual and Multicultural Development*, *45*, 1988–2004.

Siregar, I. R. (2022). The effect of note-taking strategy on listening mastery at the Grade XI Students of SMA Islam Terpadu Darul Hasan Padangsidimpuan (Unpublished doctoral dissertation). State Institute for Islamic Studies Padangsidimpuan.

Ulfani, S. T., Sumardiyani, L., & Affini, L. N. (2023). The implementation of Cornell note-taking to improving students' reading comprehension. *Jurnal Pendidikan dan Sastra Inggris*, *3*(3), 50–59.

Vuogan, A., & Li, S. (2023). A systematic review of meta-analyses in second language research: Current practices, issues, and recommendations. *Applied Linguistics Review*, *15*(4), 1621–1644.

Zhang, S., & Zhang, X. (2022). The relationship between vocabulary knowledge and L2 reading/listening comprehension: A meta-analysis. *Language Teaching Research*, *26*, 696–725.

CHAPTER 8

Learning through conducting meta-analyses

Supervisor and student reflections

Stuart Webb & Su Kyung Kim
University of Western Ontario | Korea University

Meta-analyses can appear daunting to both beginning and established researchers because they require a comprehensive understanding of the research literature and involve analytical approaches that differ from those used in most empirical studies. As a result, research students and their supervisors may hesitate to include meta-analyses in graduate research. However, conducting a meta-analysis offers several important benefits for research students. The process encourages a deeper understanding of research designs in the field and helps students critically evaluate existing methodologies and findings. It also enables them to identify research gaps that can inform future empirical studies. In addition, engaging with meta-analytic techniques helps students develop advanced statistical and analytical skills. Drawing on our experiences conducting student-led meta-analyses, this chapter discusses the challenges and benefits encountered by both students and supervisors.

Keywords: meta-analysis, PhD supervision, PhD research, MA research, vocabulary

In recent years, our research team has conducted several meta-analyses (e.g., Uchihara et al., 2019; Webb et al., 2020, 2023; Yanagisawa et al., 2020), some of which have contributed to the requirements for student degrees (e.g., Jin & Webb, 2024; Kim & Webb, 2022; Sutton & Webb, 2026; Yanagisawa & Webb, 2021; 2022). There are good reasons for students to conduct a meta-analysis as the main component of a master's degree or as the initial study within a doctorate. First, meta-analyses require the development of comprehensive knowledge of a topic that should be a target learning outcome in most research-intensive programs. Moreover, the results of the meta-analyses may reveal the most useful research questions for subsequent studies about the topic. However, meta-analyses can appear

https://doi.org/10.1075/rmal.17.08web
Available under the CC BY-NC-ND 4.0 license.© 2026 John Benjamins Publishing Company

daunting to both beginning and established researchers. They require strong and comprehensive understanding of the research literature. Moreover, they include different types of analyses from most studies, thus requiring an additional aspect of learning. For these reasons, research students may be less likely to conduct meta-analyses within their studies.

When deciding on whether the best option for students is to conduct a meta-analysis, there are several things for supervisors to consider. First, the broad scope of the search for studies that meet inclusion and exclusion criteria in a meta-analysis and the amount of detailed knowledge required of each study involve a large amount of independent work on the part of the student. This means that it might be more difficult to provide the same degree of oversight in meta-analyses than in empirical studies. For example, the database search in Kim and Webb (2022) elicited 7,246 titles and abstracts of studies that needed to be screened. Subsequently 69 studies were assessed for eligibility. Moreover, it is highly likely that there will be some ambiguity with interpretation of the details across these studies. Because studies are not designed and written to conform with each other, it is difficult to anticipate where ambiguity will occur. This means that transparency in screening and coding studies is particularly important and regular communication between students and supervisors is essential to ensure that both parties are in agreement of the process that is being followed. Therefore, it is extremely important that supervisors and students understand the challenges that might occur at the different stages of the meta-analysis, and whenever there is uncertainty, to meet and discuss the advantages and disadvantages of possible options.

Before starting a meta-analysis, graduate students may have mixed feelings. First, students may understand that conducting a meta-analysis can serve as a useful guideline for reading and interpreting prior studies about their chosen topic. This may help them to reduce their anxiety about research. Second, students may see meta-analysis as an efficient way to analyze previous research, a necessary step for their studies, while also revealing opportunities to find additional research topics beyond their dissertation. However, students may also feel burdened by the need to understand meta-analyses, which differ from more common forms of experimental research. Lastly, students will likely find the requirement of learning meta-analysis methods to be an additional challenge. This may require students to take courses — either online or offline — to acquire the necessary skills for conducting meta-analyses. While this training is valuable and necessary, it may be an added financial burden which can also add anxiety to graduate studies.

The reflections presented in this chapter are informed by our experiences conducting several systematic research syntheses using meta-analytic methods. In these studies, we adopted established procedures for research synthesis, including systematic literature searches across multiple databases, the application of explicit

inclusion and exclusion criteria, structured coding of study characteristics, and quantitative aggregation of effect sizes. These procedures were guided by widely used methodological frameworks for systematic reviews and meta-analyses (e.g., Borenstein et al., 2009; Higgins & Green, 2008; Moher et al., 2009). Drawing on these experiences, the chapter discusses the methodological and ethical challenges that supervisors and graduate students may encounter when conducting meta-analyses within research degree programs.

Methodological and ethical dilemmas

Balancing comprehensiveness and precision in defining and selecting moderators, coding their subgroups, and addressing issues of publication bias remain key methodological and ethical dilemmas for research students conducting meta-analyses. These dilemmas are not only methodological but also ethical in nature because they involve decisions about fairness in study inclusion, transparency in analytic choices, and responsibility toward both the research literature and the well-being of student researchers. These challenges also resonate with broader methodological developments in meta-analytic research within Applied Linguistics (see Plonsky, Hu, & Zhao, Chapter 3).

Scope

As with most types of research, an initial challenge can be deciding on the scope of the study. This is particularly important with meta-analyses because when the scope increases, the search demands may become unmanageable. For example, it might be tempting in meta-analyses of L2 learning to also include studies of L1 learning and code L1-L2 as a moderator variable to look at how that affects gains. That might provide an interesting comparison. However, if the focus of the study is frequently researched (e.g., incidental vocabulary learning, spacing effects), the scope of the search and the extent to which the data could be analysed and included within a single study may go beyond what is manageable. Students (and supervisor!) may often be motivated to take on a wider scope than is manageable. Thus, it is important to ensure that the scope of the meta-analysis can provide an achievable outcome.

Inclusion of non-WEIRD (Western, Educated, Industrialized, Rich, and Democratic) studies

One of the more challenging aspects of the literature search is the ethical question of whether non-peer reviewed research and studies from lesser known sources should be included in the meta-analysis. Because a comprehensive literature search may identify a large number of studies from a wide range of sources, supervisors are unlikely to be familiar with all of the studies that meet the inclusion and exclusion criteria. A question that may then arise is how to ensure that the included studies meet expected quality standards? A meta-analysis may yield incorrect or misleading results if the methodologies of the included studies do not provide sufficient information for transparent interpretation of learning conditions or the methodological rigour to effectively answer the research questions (Boers et al., 2021). This may lead to the ethical dilemma of how to conduct a comprehensive search without disadvantaging research such as conference proceedings, master's and doctoral dissertations, and articles from non-WEIRD contexts that might be published outside of more prestigious publications. This issue raises an ethical consideration regarding fairness in knowledge representation. Excluding studies from less visible outlets or non-WEIRD contexts may systematically bias the evidence base and limit the generalizability of meta-analytic conclusions.

Coding

When conducting a meta-analysis, research students often face methodological challenges, particularly in coding prior studies' methodologies. In the initial stages of coding studies, it is difficult to anticipate all moderator variables. The variables that have received the most attention in the literature are easy to anticipate and are likely to be clear. For example, in Kim and Webb's (2022) coding of studies of repeated practice, the moderators *provision of feedback* (present, absent) and *feedback timing* (immediate, delayed) were anticipated and transparent. Similarly, because there are many studies investigating the effects of spaced practice on vocabulary learning, it was easy to anticipate that *vocabulary* would be one category for the moderator *learning target*. However, it was less clear whether *vocabulary* would include both studies of single word learning and multiword learning or whether these types of studies should be coded as separate categories. It was also unclear whether less frequently investigated learning targets such as *grammar*, *pronunciation*, and *processing* might be included as categories. In addition to methodological consistency, coding decisions also have an ethical dimension

because inconsistent interpretation of study characteristics may influence the transparency and credibility of the synthesis.

Students may be inclined to include as many potential moderators as possible to demonstrate a thorough examination of variables. However, some moderators may not have sufficient sample sizes (k) in the dataset. This can result in small subgroup sizes for each moderator, making it difficult to draw reliable conclusions. However, due to findings from earlier studies, students may hesitate to exclude certain moderators, even when their inclusion is not statistically justified. For example, individual studies often have varying conditions (e.g., stimuli, target items or features, the number of repetitions), leading students to feel pressured to incorporate all possible conditions while simultaneously striving for a more focused analysis.

Creating labels for moderator variables

One major issue when coding studies can be creating understandable and widely acceptable labels for moderators. Because studies are designed for different purposes and written by different researchers, there may be little conformity in the terms that are used to describe learner, treatment, and methodological variables. Moreover, methodologies may often lack sufficient detail, making it difficult to interpret the variables of interest. For instance, defining and coding the type of knowledge measured, participant proficiency level, and test type can be challenging, because of variation in the terms used, while aspects of the design such as treatment and test duration may not be transparent or always included. This will likely be a major issue for students, because they are the ones who are trying to interpret and code the studies. However, it is also something that supervisors need to anticipate because differences in interpretation of terminology and definitions will lead to inconsistencies in the coding process, which in turn may impact the findings.

Another issue may be coding a study that combines multiple conditions. For example, when a study includes both comprehension-based (e.g., reading or listening) and production-based (e.g., sentence writing, picture description) learning conditions, students may categorize it as *mixed*. However, this broad classification can make result interpretation difficult, especially when the *mixed* subgroup has a larger sample size than other subgroups under a moderator. Such inconsistencies can obscure meaningful distinctions between different learning conditions and reduce the clarity of the analysis.

Statistical knowledge

Because meta-analyses are less common than intervention studies, it may be challenging for students (and supervisors!) to gain the methodological and statistical expertise required. Methodologically, meta-analyses require systematic literature searches and screening procedures (e.g., constructing comprehensive search strategies, applying rigorous inclusion and exclusion criteria, and following systematic review protocols to ensure transparency and replicability). They also involve the careful identification and operationalization of study characteristics (e.g., coding and data extraction) and the critical evaluation of study quality (e.g., randomization, control conditions, reporting completeness). Statistically, students conducting meta-analyses must be able to compute and interpret effect sizes from diverse reported data, understand model selection (fixed- or random-effects), and evaluate heterogeneity (e.g., Q statistic, I^2) and potential publication bias using advanced techniques such as moderator analyses (e.g., meta-regression, subgroup analysis) and funnel plots (e.g., Egger's regression). Unlike single empirical studies that analyze researcher-generated data within a single design, meta-analyses synthesize results across multiple studies, requiring not only the ability to manage methodological diversity and interpret cumulative evidence across the field but also statistical proficiency in synthesizing heterogeneous data. There may not be available courses on how to conduct a meta-analysis nor how to analyse the data. Indeed, this was the situation in our university context. Therefore, it may be necessary to look elsewhere to develop the skills necessary to conduct meta-analyses.

Writing up the meta-analysis

Writing up a meta-analysis presents challenges to both supervisors and students. Writing for publication may be the most difficult aspect of graduate studies for all students regardless of their English language background. Learning how to write a meta-analysis article presents students with many challenges because along with the focus on the primary research question, the article also needs to present explanation, background, and discussion of the many moderator variables. Describing all aspects of the study with sufficient detail and keeping within the word limit can be difficult. For example, how much detail should be provided about the different studies that are discussed in the Introduction and Background sections? How much detail is necessary in the Methodology? To what extent should the findings be discussed in the Results section? What can be included in the Discussion outside of the results? Finally, how can all of this information be included in a manuscript with a 8,500 or 10,000 word limit? Because writing a manuscript can be

stressful for students, supervisors are tasked with providing both the support and critical feedback necessary for a positive learning experience. For supervisors this may include the following challenges. First, the student's narrative of the manuscript may differ from how the supervisor may have framed it. This should be fine because there are likely to be several ways to successfully write a manuscript. It should also be considered a positive outcome because it allows students to learn first from the feedback of their supervisor and then subsequently from the feedback of examiners and reviewers. However, a challenge for supervisors is ensuring that students learn through the writing process rather than having their writing difficulties solved by the supervisor. This can be a time-consuming process that involves going through many drafts of the manuscript. In the end, however, it allows students to learn through the process and hopefully develop the skills necessary to write improved drafts of subsequent manuscripts.

Student well-being

Perhaps, the most important ethical issue for supervisors and students to consider before conducting a meta-analysis is the well-being of student researchers. Supervisors therefore have an ethical responsibility to consider how the demands of conducting a meta-analysis may affect students' workload, stress levels, and sense of isolation during the research process. Of course, this is something that should be considered in relation to any student research project. However, there might be a couple of issues that are more relevant with meta-analyses than other types of research. For example, because meta-analyses are less common than intervention studies, students might be less able to draw on and learn from the resources within the university, faculty members, and their peers. At our university, while there are many courses offered on different types of statistical analyses, there was not one that included meta-analysis. Moreover, although our research team has conducted quite a few meta-analyses, at the time each one was undertaken, most were unique to the studies being conducted by the rest of research team. In addition, the literature search and coding of studies in meta-analyses tends to involve a very large number of hours reading over a period of months. This may prompt a greater amount of time on their own and lead to feelings of isolation and loneliness for students. Finally, the challenges of learning a new form of statistical analyses may cause anxiety if there are few peers around to learn from.

Course of action

Scope

For students, the scope of the study may present unforeseen challenges. This is because closely related topics may emerge when screening studies during the search process. One example of this was how to address frequency effects when screening studies for a meta-analysis of meaning-focused L2 vocabulary learning. Because many studies of meaning-focused L2 vocabulary learning examine the effects of frequency of occurrence of target words, it made sense to code for frequency because it was a theoretically motivated variable. A decision then needed to be made about whether to expand the focus of the meta-analysis to include frequency effects or examine frequency effects on its own. In this case, we felt it would be better to do two meta-analyses, one focused on frequency effects in L2 vocabulary learning (Uchihara et al., 2019) and the other on meaning-focused L2 vocabulary learning (Webb et al., 2023). Including both within a single article would have limited the detail that could have been covered in each due to word limit restrictions. Similarly, in a meta-analysis of L2 learning through viewing meaning-focused audiovisual input, Sutton and Webb (2026) found that many studies included closely related learning conditions such as learning through viewing L2 meaning-focused audiovisual input with captions or subtitles. A decision then had to be made about whether it would be useful to expand the scope of the study to also include one or both of these conditions. In this case, we decided that expanding the scope would present unmanageable demands on screening and coding of the studies because the topic of viewing L2 meaning-focused audiovisual input with captions or subtitles was very large. Moreover, the findings for the original topic, learning through viewing L2 meaning-focused audiovisual input, might be less transparent and meaningful when combined with studies of captions and subtitles.

Inclusion of non-WEIRD studies

We have taken several steps to ensure that the included studies met expected quality standards without disadvantaging research published outside of prestigious outlets. First, we have included doctoral dissertations within our search (e.g., Kim & Webb, 2022; Uchihara et al., 2019; Yanagisawa et al., 2020). The advantage of including PhD dissertations within the search is that it reduces the chance of publication bias affecting the results. Publication bias refers to the increased likelihood that results of studies published in journals will show statistically significant findings in comparison to those in dissertations. Second, we never limited the

search to only prestigious journals. There are several reasons why we believe this is useful. These include the fact that no matter the status, all peer-reviewed journals are likely to include some articles that have methodological issues. Similarly, because a journal is less well known does not mean that research quality is poor. In some contexts, researchers may be encouraged to submit their research to local journals, may have a need to have their research published more quickly, or may lack the confidence to believe that publication will occur in more prestigious outlets. By including a wide variety of publication outlets in the search, researchers from non-WEIRD contexts are less likely to be disadvantaged.

The third step that we have taken to ensure a comprehensive search that follows ethical guidelines is to carefully check that the research meets the inclusion and exclusion criteria, and when there is a lack of clarity, we meet to discuss the study and how to move forward with it. The fourth step we have taken is running additional analyses to see how the inclusion of any studies influences research findings. This provides greater transparency for readers because it allows them to see how the inclusion of certain studies may affect findings. For example, we reported sensitivity and influence analyses — including publication bias tests (fail-safe N, trim-and-fill, funnel plots) — and examined variance inflation factors (VIFs) to assess multicollinearity in the meta-regression models (e.g., Kim & Webb, 2022; Webb et al., 2023). These checks demonstrated that our results were not driven by certain studies.

Coding

Differences in definitions can lead to inconsistencies in the coding process. One problem that we faced when coding variables in our meta-analysis on spaced practice (Kim & Webb, 2022) was how to define the moderator variable *retention interval* because researchers use different definitions. Initially, we defined it as the interval between the end of practicing the target feature and testing. For example, if the last day of practice occurred on July 1 and the delayed posttest occurred on July 15, then the retention interval was 14 days. This definition was based on studies of meaning-focused and intentional vocabulary learning (e.g., Webb et al., 2023; Webb et al., 2020). However, as noted by one reviewer, Suzuki (2017) reported that in studies of spaced learning, when posttests are administered to participants at multiple intervals (e.g., immediate posttest, delayed posttest 1, delayed posttest 2), the immediate posttest and first delayed posttest may serve as additional practice sessions. Thus, if the schedule for testing was immediate posttest on July 1, delayed posttest 1 on July 8, and delayed posttest 2 on July 15, the retention interval should be seven days because learning can occur through the earlier posttests. In this case, while there was justification for both approaches,

we followed the recommendation of the reviewer to incorporate Suzuki's (2017) interpretation, which was specific to studies of repeated practice.

Creating labels for moderator variables

Deciding on the labels for moderator variables may present challenges. Because of variation in the terms used across studies and differences in the ways in which variables are operationalized in the research literature, choosing labels for variables that are transparent and justified can be difficult. For instance, defining and coding moderators such as *test type*, *measurement*, *number of repetitions* and *frequency of practice* were challenging for Kim and Webb (2022), because researchers used different labels and definitions. As a result, we struggled to establish widely persuasive labels and definitions that could accommodate diverse perspectives. For example, two moderator variables were labeled *frequency of practice* and *number of sessions*. The variables were justified within the literature, because *frequency of practice* can refer to both within-session and between-session practice, whereas *number of sessions* refers to the individual session in which learning took place. However, if not well-acquainted with these definitions, the labels may seem to refer to the same variable, and so detailed and transparent definitions were required to reduce the potential for misinterpretation.

Statistical knowledge

Developing the methodological and statistical expertise required to conduct meta-analyses is an ongoing challenge. Our research team initially learned meta-analyses through self-study. Borenstein et al.'s (2009) *Introduction to meta-analysis* was an essential resource, as was the *Comprehensive Meta-Analysis* software used to analyse data. One of the team members also enrolled in the course *Comprehensive Meta-Analysis* through meta-analysis.com, which further improved understanding of meta-analysis as well as the software used for analysis. Working as a team and supporting each other has been an important component of the success of our meta-analyses and this is particularly true of developing statistical knowledge.

Writing the meta-analysis

Finding useful models of published meta-analyses may be the most useful starting point for students setting out to write up a meta-analysis. Meta-analyses published in good journals with the same or similar word limits as the target journal provide students with models of how to organize the different sections of the article and

raise awareness of which information is required and the level of detail necessary for each section. Initially, there were few meta-analyses closely linked to our topics so those that we found provided a starting point to follow for how we might write up our meta-analyses (e.g., de Vos et al., 2018; Swanborn & de Glopper, 1999). L1 meta-analyses also provided useful models for how a topic might be discussed. For example, for our L2 meta-analysis of spaced practice, we drew on Donovan and Radosevich (1999) and Cepeda et al.'s (2006) meta-analyses of the effects of spaced retrieval on L1 learning. It is important to note that no individual study is likely to provide a perfect model of how to write up a new study. Each study is written according to different parameters such as the journal constraints and the feedback of reviewers and editors. Moreover, if the authors of an earlier published work were to write the study again, it is likely to be different because knowledge and reporting guidelines tend to change over time. However, several different studies may help reveal how the different parts of each section of an article may be written and this may ease the writing burden to some degree.

If meeting the word limit of a target publication is challenging, including some of the findings and analyses in supplementary materials is a useful strategy. In Kim and Webb (2022), ten appendices covering 68 pages were included in supplementary materials. These appendices covered information that was necessary for replication such as the PRISMA flow diagram, coding scheme, category criteria, and details of included studies in the meta-analysis. Appendices also included important analyses that were challenging to include due to space limitations such as publication bias analyses, and moderator analyses for each posttest under each category. By including this information in supplementary materials, the narrative of the manuscript may read better from start to finish and readers can access other necessary information to evaluate the research in the supplementary materials if they are interested.

Student well-being

To help ensure the well-being of students, it is most important for supervisors to raise awareness of the challenges that might be faced while conducting a meta-analysis. Students should understand the extent to which they will be able to get support from the university, their supervisor, faculty members, and peers. They should also be aware of the time-intensive nature of conducting meta-analyses, the large amount of critical reading involved, and the potential for this to feel isolating. Because there should typically be multiple options for student research, discussing the advantages and disadvantages of each option before moving forward with a study may allow students to select the best option for their needs.

In our research team, we initially conducted four meta-analyses as a three-person team consisting of a supervisor and two students (Uchihara et al., 2019; Webb et al., 2020, 2023; Yanagisawa et al., 2020). Working as a team helped to develop knowledge of how to conduct meta-analyses, clarify any misconceptions that occurred along the way, and build on the knowledge gained through conducting each study. Support could be provided between students, from supervisor to student and from student to supervisor. Moreover, the knowledge gained through conducting the initial meta-analyses helped support new meta-analyses conducted by subsequent students (e.g., Jin & Webb, 2024; Kim & Webb, 2022). By working together as a team and sharing knowledge, students may find meta-analysis more collaborative and less isolating, which may enhance student well-being. Similar supervisory dynamics are described by Chalmers and Hamilton (Chapter 15), who discuss collaboration between doctoral researchers and supervisors when conducting systematic reviews.

Alternative courses of action

Scope

In the previous examples related to scope, the most appropriate course of action was to limit the scope of the research. This allowed us to provide a more detailed and transparent analysis of individual topics. However, at times it may also be useful to expand the scope of the research. For example, Kim and Webb (2022) included a more comprehensive scope of studies in their meta-analysis of studies of repeated practice than Uchihara et al. (2019), Webb et al. (2023), and Sutton and Webb (2026). Kim and Webb needed to consider whether to include studies of the different schedules for repeated practice (spaced vs massed, shorter vs longer spacing, expanding vs equal spacing) in one article, or to examine each on its own. In this case, Kim and Webb decided to include all three types of spacing in a single study to provide a more comprehensive account of the effects of repeated practice. This allowed readers to gain greater insight into how the different types of studies of repeated practice contribute to L2 learning within one article. The disadvantage of this approach was that by including multiple meta-analyses within a single article, each topic was condensed, perhaps making the article less reader friendly.

Inclusion of non-WEIRD studies

Earlier we discussed the steps that were taken to ensure quality standards are met without disadvantaging non-WEIRD studies. However, mitigating publication bias in meta-analyses also involves thorough study identification and the use of bias-detection tools. As a general strategy, students can conduct exhaustive literature searches across multiple databases and include *gray literature* (e.g. dissertations, conference proceedings, and other unpublished studies) to capture relevant findings beyond the published domain (Higgins & Green, 2008). This inclusive approach helps ensure that studies with nonsignificant or unfavorable results — which might otherwise remain unpublished — are not systematically overlooked (Higgins & Green, 2008). Clear inclusion criteria for the review can guard against included studies based on their outcomes and promote a transparent, unbiased selection process. To describe the study inclusion criteria and the number of included and excluded references in the process, students can provide a PRISMA flow diagram (Moher et al., 2009). Students can also employ specific statistical techniques to detect and address publication bias in the gathered data (Borenstein et al., 2009). Publication bias could be probed visually and analytically (funnel plots, Egger's test) and, where appropriate, adjusted (trim-and-fill or selection models), alongside comparisons of published vs. gray literature (e.g., Egger et al., 1997; Stanley & Doucouliagos, 2014). By combining comprehensive search efforts with these analytical tools, students can greatly reduce the impact of publication bias and strengthen the credibility of their meta-analytic conclusions. Additionally, students can remove the flagged outlier (e.g., Zhao & Aryadoust, 2024). Zhao and Aryadoust employed influence diagnostics (e.g., externally studentized residuals) to detect overly influential studies and then conducted sensitivity analyses comparing models with and without the outlier.

Students can pilot the eligibility criteria on a subset of studies (e.g., 10–20%), have two coders screen that subset independently, and reconcile disagreements to surface ambiguous cases. The team can then revise the coding dataset and report inter-rater agreement (e.g., κ or percent agreement, e.g, Norouzian, 2021) before proceeding to full screening. The team can then run follow-up robustness checks (e.g., studentized residuals, outlier removal, publication bias tests). These steps help ensure that inclusion/exclusion decisions and analytic choice do not excessively influence the overall conclusions. Such procedures reflect the field's growing emphasis on transparency and rigor in meta-analyses and enhance the credibility of the findings.

Coding

As discussed previously, it is important to ensure that any moderator variable coded in the meta-analysis is clearly defined and grounded in prior studies, providing a strong theoretical justification for its inclusion. The data from moderator variables is unlikely to be consistent and so supervisors and students may need to decide how to treat variables that do not quite conform to expected parameters. For example, it is important to decide on how to treat moderator variables that are less frequent within the data. If a moderator subgroup is represented by only a few studies, it may be more appropriate to synthesize those results narratively rather than statistically, particularly when the moderator is theoretically important but the quantitative data are too sparse for robust analysis. Similarly, when included studies are coded under a broad *mixed* category, students can describe the individual study designs in detail to illuminate the heterogeneity within that grouping instead of treating *mixed* as a homogeneous label.

Creating labels for moderator variables

The approach we discussed previously involved providing detailed and transparent definitions that were grounded in the literature. However, this might not always be possible. Because there tends to be variation in how activities used for learning are operationalized across studies, it may be difficult to classify activities according to well-established labels. This is because often a treatment may involve multiple activities that contribute to learning. For example, Bloom and Shuell (1981) included both multiple-choice and fill-in-the blank activities within their learning condition. Because these activities are quite different and involve different types of psychological conditions that contribute to learning (multiple-choice contributes to learning through receptive recognition while fill-in-the-blanks contributes to learning through form recall), it was not easy to decide on how to fit this under a transparent category label that also fit other activity types. In this case, we created broader categories (comprehension activities, production activities, combined activities that include both comprehension and production exercises) that allowed transparent interpretation of coding. An advantage of using a broader more inclusive label is that it may increase the number of studies and effect sizes within a category. The disadvantage of this approach is that the results are less precise because the category includes different activities.

Statistical knowledge

Earlier we discussed the different ways in which the statistical knowledge required for a meta-analysis might be gained. Here we discuss how gaining a more thorough understanding of the statistical methods required to conduct meta-analysis support the research process. Applying rigorous statistical methods improves the comprehensiveness, transparency, and credibility of meta-analyses. It is important to attend to selecting and transforming effect sizes appropriately (e.g., Hedges' *g* with a small-sample correction, Fisher's *z* for *r*, log odds for proportions) (Borenstein et al., 2009; Cooper, 2017; Hedges & Olkin, 1985; Lipsey & Wilson, 2001; Morris, 2008). When a single study yields multiple outcomes, time points, or contrasts, model that dependence — preferably via multilevel/multivariate meta-analysis or robust variance estimation with small-sample adjustments — rather than averaging effects ad hoc (Cheung, 2014; Hedges et al., 2010; Van den Noortgate et al., 2013). Heterogeneity could be quantified and interpreted with Q, τ^2, I^2 and, crucially, prediction intervals when k is small (Kim & Webb, 2022). Meta-regression requires careful specification (pre-registered moderators, centering, checks for multicollinearity via VIFs) and defensible inference; report model comparison and the proportion of heterogeneity explained (Knapp & Hartung, 2003; Riley et al., 2011; Thompson & Higgins, 2002; Veroniki et al., 2016). Robustness could be demonstrated with influence and sensitivity analyses (leave-one-out, standardized residuals; alternative τ^2 estimators; alternative effect-size computations; varying assumed within-study correlations for dependent effects) (e.g., Viechtbauer & Cheung, 2010).

Writing the meta-analysis

Previously we discussed how models of published meta-analyses may reveal the easiest pathways to writing up the study. Meta-analyses tend to be relatively formulaic in the level of detail and organization of that information. This can be particularly true of Discussion sections which often sequence answers to the main research questions followed by the interpretation of the moderator variables. Because there can be quite a few moderator variables, there may be little space available for more detailed discussion. This may lessen the writing burden for students. However, it is important to note that there are different ways to write up meta-analyses and sometimes the most common model may not necessarily provide the best pathway. For example, the data examined within a meta-analysis typically provides the most transparent findings for the primary research question, because the greatest amount of data is used to answer that question. In contrast, moderator variables typically include only a portion of the total data set to

answer the extent to which they influence findings. This means that the degree to which the effects of the different categories for a moderator variable are revealed by the analyses is reduced. Therefore, it may be useful for supervisors to suggest that their students include greater detail in interpreting the findings for the primary research questions. While it is still important to provide some interpretation of the moderator analyses, there may be value in devoting slightly less space to this part of the discussion particularly when sample sizes are small.

The findings for the primary research questions should be interpreted carefully, in accordance with domain-specific guidance and benchmarks, to aid reader comprehension. Many meta-analyses in L2 research align interpretations with SLA-specific guidance and benchmarks (Norouzian & Bui, 2024; Plonsky, 2013; Plonsky & Brown, 2015; Plonsky & Oswald, 2014). In our meta-analysis (Kim & Webb, 2022), the theories and ideas of spaced practice derive from psychology. Accordingly, we aligned our effect-size interpretations with both SLA-specific benchmarks (e.g., Plonsky & Oswald, 2014) and psychology-specific benchmarks (Schäfer & Schwarz, 2019). For example, an effect size of 0.50 is typically interpreted as medium under the SLA benchmarks but as large in psychology research.

Student well-being

Previously we discussed the support that supervisors and research team members can provide for students conducting meta-analyses. However, support does not have to be internal. Support may be gained from students, faculty members, and researchers outside of the immediate faculty and university contexts. Although universities aim to provide a sufficient network of internal support for students, there is no reason why support cannot be developed across universities. In fact, it makes sense that support systems are created outside of a university because there is great value for students in expanding their research networks beyond a program and university. One example of how this can be done is the Interuniversity PhD Network (iUP) which was developed to provide support, learning, and collaboration opportunities between PhD students in Applied Linguistics programs at University College London, University of Hawaii at Manoa, and the University of Western Ontario (Iwaizumi et al., 2025). Support might also be provided through contacting individual researchers at other institutions. Considering options for support outside of the university may provide a pathway for greater student support and well-being.

Conclusion

The aim of this chapter has been to shed light on some of the benefits and challenges of conducting a meta-analysis within a research degree. Our research team has had some success in undertaking meta-analytic studies in MA (Sutton & Webb, 2026) and PhD degrees (e.g., Jin & Webb, 2024; Kim & Webb, 2022; Yanagisawa & Webb, 2021, 2022). However, there was a great deal of learning for both students and the supervisor in each of these studies. We have discussed where some of the challenges for supervisors and students may occur during a meta-analysis. Research degrees are opportunities for learning, and we have learned a great deal through conducting meta-analyses.

There are good reasons to include a meta-analysis as a study within a research degree. First, students completing a meta-analysis are likely to achieve targeted learning outcomes. They should develop comprehensive knowledge of a topic, gain valuable skills in critical thinking and can develop methodological and statistical expertise. However, it is important that both supervisors and students are aware of the benefits and challenges that they may experience through the process of completing a meta-analysis.

Because meta-analytic studies involve a large amount of independent work on the part of students, regular communication between students and supervisors is extremely important. Moreover, because there is likely to be ambiguity across the literature in definitions, labels, and operationalization of constructs, there are likely to be different pathways that could be taken during coding. Thus, meeting and discussing the different options for coding is critical. The review process may allow students to develop clearer definitions of key terms and moderator labels, which improves the consistency and accuracy of their coding. Additionally, by analyzing many studies, students gain a deeper understanding of research designs within their field, helping them critically evaluate methodologies and findings. This broader perspective enables them to identify research gaps in previous studies, which can serve as a foundation for future empirical research, an especially valuable advantage for MA and PhD students preparing for future studies. Furthermore, engaging with meta-analytic techniques helps students acquire advanced statistical knowledge, which not only enhances their ability to assess overall effectiveness in their research area but also equips them with transferable analytical skills applicable to other fields. Together with other chapters in this volume that reflect on different forms of research synthesis — including qualitative synthesis (Kato, Section 5 overview), scoping reviews (Liu, Chapter 11), and systematic reviews (Oxley, Chapter 13) — this chapter highlights the importance of reflective methodological decision-making when conducting research syntheses in Applied Linguistics.

Questions for discussion

1. What are the different sources of support for students conducting meta-analyses in your university context?
2. Which types of support are most needed for students conducting meta-analyses in your university context?
3. What are the most effective ways for supervisors to help their students navigate the process of conducting meta-analyses?
4. What are some topics that should be discussed between supervisors and students before undertaking a meta-analysis as part of a research-based degree?

Further reading

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Wiley & Sons.

This book is the most comprehensive work on meta-analysis. It provides clear explanations of effect sizes, heterogeneity, moderator analysis, and publication bias, along with conceptual illustrations. We recommend this book because it offers a solid foundation in both the logic and the practice of meta-analysis, suitable for students new to quantitative synthesis.

Cooper, H. (2017). *Research synthesis and meta-analysis* (5th ed.). Sage.

This book emphasizes the process of research synthesis from formulating research questions and conducting systematic search to coding and reporting results. It is useful for students to understand balanced treatments of conceptual procedural and statistical aspects of synthesis.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Sage.

This book provides step-by-step guidance on conducting meta-analyses with real data examples. We recommend this book because it helps bridge the gap between theory and application and planning to carry out meta-analyses.

Norouzian, R. (2021). Interrater reliability in second language meta-analyses. *Studies in Second Language Acquisition*, *43*(4), 896–915.

This article provides a comprehensive discussion of reliability indices, interpretation, and reporting practices specific to L2 studies. We recommend it because it helps strengthen the methodological rigor of meta-analyses and promote transparent and replicable coding procedures.

Norouzian, R., & Bui, G. (2024). Meta-analysis of second language research with complex research designs. *Studies in Second Language Acquisition*, *46*(1), 251–276.

This article provides the methodological sophistication of meta-analyses in L2 research by addressing how to synthesize studies with complex designs. We recommend this paper for researchers seeking to understand strategies for handling dependent effects sizes and multilevel structures using modern statistical approaches.

Norouzian, R., Jin, Z., & Webb, S. (2025). Increasing meta-analytic quality: A multivariate multilevel meta-analysis of note taking through exposure to L2 input. *Modern Language Journal*, *109*(1), 171–193.

This article shows how a classic meta-analysis approach can be improved through the use of a multivariate, multilevel meta-analysis approach. We recommend this paper because it provides clear models of each approach while also revealing the steps that can be taken to move from a classic to a multivariate, multilevel meta-analysis approach.

AI use declaration statement

The authors confirm that no AI tools were used in the preparation, writing, formatting, and editing of this work.

References

Bloom, K. C., & Shuell, T. J. (1981). Effects of massed and distributed practice on the learning and retention of second-language vocabulary. *Journal of Educational Research*, *74*(4), 245–248. Retrieved from

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin*, *132*(3), 354–380. Retrieved from

Cheung, M. W.-L. (2014). Modeling dependent effect sizes with three-level meta-analyses: A structural equation modeling approach. *Psychological Methods*, *19*(2), 211–229.

De Vos, J. F., Schriefers, H., Nivard, M. G., & Lemhöfer, K. (2018). A meta-analysis and meta-regression of incidental second language word learning from spoken input. *Language Learning*, *68*, 906–941.

Donovan, J. J., & Radosevich, D. J. (1999). A Meta-Analytic Review of the Distribution of Practice Effect: Now You See It, Now You Don't. *Journal of Applied Psychology*, *84*, 795–805.

Egger, M., Davey Smith, G., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ*, *315*(7109), 629–634.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Academic Press.

Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect sizes. *Research Synthesis Methods*, *1*(1), 39–65.

Higgins, J., & Green, S. (2008). *Cochrane handbook for systematic reviews of interventions*. Cochrane.

Iwaizumi, E., Nuesser, M., & Kim, J. (2025). Writing Fridays: Connecting doctoral students across continents. In R. Cayley, F. Coll, & A. Newman (Eds.), *Writing together: Building social writing opportunities for graduate students* (pp. 82–88). University of Michigan Press.

Jin, Z., & Webb, S. (2024). The effectiveness of notetaking through exposure to L2 input: A meta-analysis. *Studies in Second Language Acquisition*, *46*(2), 404–426.

Kim, S. K., & Webb, S. (2022). The effects of spaced practice on second language learning: A meta-analysis. *Language Learning*, *72*(1), 269–319.

Knapp, G., & Hartung, J. (2003). Improved tests for a random effects meta-regression with a single covariate. *Statistics in Medicine*, *22*(17), 2693–2710.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & the PRISMA Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLOS Medicine*, *6*(7), e1000097.

Morris, S. B. (2008). Estimating effect sizes from pretest–posttest–control group designs. *Organizational Research Methods*, *11*(2), 364–386.

Plonsky, L. (2013). Study quality in SLA: A meta-analysis. *Studies in Second Language Acquisition*, *35*(4), 655–687.

Plonsky, L., & Brown, D. (2015). Domain definition and search techniques in meta-analyses of L2 research (or why 18 meta-analyses of feedback have different results). *Second Language Research*, *31*(2), 267–278.

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning*, *64*(4), 878–912.

Riley, R. D., Higgins, J. P. T., & Deeks, J. J. (2011). Interpretation of random effects meta-analyses. *BMJ*, *342*, d549.

Schäfer, T., & Schwarz, M. A. (2019). The meaningfulness of effect sizes in psychological research: Differences between sub-disciplines and the impact of potential biases. *Frontiers in Psychology*, *10*, 813.

Stanley, T. D., & Doucouliagos, H. (2014). Meta-regression approximations to reduce publication selection bias. *Research Synthesis Methods*, *5*(1), 60–78.

Sutton, D., & Webb, S. (2026). The effects of audiovisual input on second language learning: A meta-analysis. *Studies in Second Language Acquisition*.

Suzuki, Y. (2017). The optimal distribution of practice for the acquisition of L2 morphology: A conceptual replication and extension. *Language Learning*, *67*(3), 512–545.

Swanborn, M. S. L., & de Glopper, K. (1999). Incidental word learning while reading: A meta-analysis. *Review of Educational Research*, *69*, 261–285.

Thompson, S. G., & Higgins, J. P. T. (2002). How should meta-regression analyses be undertaken and interpreted? *Statistics in Medicine*, *21*(11), 1559–1573.

Uchihara, T., Webb, S., & Yanagisawa, A. (2019). The effects of repetition on incidental vocabulary learning: A meta-analysis of correlational studies. *Language Learning*, *69*(3), 559–599.

Van den Noortgate, W., López-López, J. A., Marín-Martínez, F., & Sánchez-Meca, J. (2013). Three-level meta-analysis of dependent effect sizes. *Behavior Research Methods*, *45*(2), 576–594.

Veroniki, A. A., Jackson, D., Viechtbauer, W., Bender, R., Bowden, J., Knapp, G., Kuss, O., Higgins, J. P. T., Langan, D., & Salanti, G. (2016). Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Research Synthesis Methods*, *7*(1), 55–79.

Viechtbauer, W., & Cheung, M. W.-L. (2010). Outlier and influence diagnostics for meta-analysis. *Research Synthesis Methods*, *1*(2), 112–125.

Webb, S., Uchihara, T., & Yanagisawa, A. (2023). How effective is incidental vocabulary learning? A meta-analysis. *Language Teaching*, *56*(2), 161–180.

Webb, S., Yanagisawa, A., & Uchihara, T. (2020). How effective are intentional vocabulary learning activities? A meta-analysis. *Modern Language Journal*, *104*(4), 715–738.

Yanagisawa, A., Webb, S., & Uchihara, T. (2020). How do different forms of glossing contribute to L2 vocabulary learning from reading? A meta-regression analysis. *Studies in Second Language Acquisition*, *42*(2), 411–438.

Yanagisawa, A., & Webb, S. (2022). Involvement Load Hypothesis Plus: Creating an improved predictive model of incidental vocabulary learning. *Studies in Second Language Acquisition*, *44*(5), 1279–1308.

Yanagisawa, A., & Webb, S. (2021). To what extent does the involvement load hypothesis predict incidental L2 vocabulary learning? A meta-analysis. *Language Learning*, *71*(2), 487–536.

Zhao, H., & Aryadoust, V.A. (2024). A meta-analysis of the reliability of second language reading comprehension assessment tools. *Studies in Second Language Acquisition*, *47*(1), 388–416.

SECTION 5

Qualitative research synthesis and practice review

Section overview

What are qualitative research synthesis, qualitative synthesis of research, and practice review?

Sho Kato
University of St Andrews

Definition

There are two types of qualitatively informed research synthesis: qualitative research synthesis (QRS) and qualitative synthesis of research (QSR). Chong et al. (2024) defined QRS as "research synthesis that employs qualitative methods to synthesise qualitative research evidence" (p. 3) and QSR as "research synthesis that employs qualitative methods to synthesise research evidence derived from both qualitative and quantitative paradigms" (p. 3). Research syntheses that adopt qualitative analysis provide researchers and teachers with in-depth and ecologically valid evidence, promoting "reflective teaching and evidence-based innovations" (Chong & Plonsky, 2021, p. 1026) by synthesising qualitative findings in naturalistic studies. Chapter 9 in this section presents a supervisor's experience of guiding two master's students through their respective research synthesis: QRS and QSR. Through the lens of the lead author, the chapter explores the methodological and ethical dilemmas that emerged before and during the QRS/QSR processes while supervising both students.

A practice review is another type of secondary research that employs primarily qualitative methods to synthesise both academic and non-academic resources. Some examples of non-academic resources include government reports, blogs, teaching materials, and practitioner magazines. Qualitatively informed research syntheses and practice review share some similarities; for instance, both focus on in-situ pedagogical interventions and aim to bridge the gap between research and practice. However, practice review is distinctive in that its sources include materials that are rich in practice-related information, making it highly relevant to language teachers and policymakers (see Chapter 1 in this book). The purpose of a practice review is to offer an overview of current practices such as pedagogies.

https://doi.org/10.1075/rmal.17.09kat

Chapter 10 shares an experience of conducting a practice review, reflectively looking back on the process of a review on translanguaging practices in Anglophone classrooms. Its reflective discussion offers insights into methodological and ethical issues for those considering conducting a practice review collaboratively.

Methodological stages

The following methodological stages are based mainly on Chong and Plonsky (2021) and the 'Method' items in the Synthesis Methods and Reporting Tool (SMART) (Chong, 2025).

1. Design research questions
   This is a vital stage in conducting a qualitatively-informed research synthesis and practice review. Their research questions should carefully be considered to elucidate certain complex phenomena (e.g., what, why, or how), but do not necessarily have to be the questions with those interrogatives. For example, Chen (2016, p.367) set for the qualitative research synthesis the following research questions:
   1. What are the characteristics of technology-supported peer-feedback activities found in the primary studies from 1990 to 2010?
   2. What are the advantages and disadvantages in technology-supported peer-feedback activities demonstrated in the primary studies?
   3. What are the technologies used in the synchronous or asynchronous types of peer interaction in the primary studies and what are the differences between synchronous or asynchronous types of peer interaction?
   4. What are the main themes emerging from the GT analysis and what is the metatheory for the synthesis?
2. Establish inclusion and exclusion criteria
   This stage involves establishing clear criteria for screening articles. Inclusion and exclusion criteria should be predicated on elements such as its research design and the focus of the domain of interest. In the case of a qualitative research synthesis, a qualitative or mixed research design should be considered. The inclusion of raw data, interpreted data, or both should also be taken into consideration. For a practice review, closer-to-practice criteria, which could be related to current issues practitioners face, should be included. The involvement of practitioners would be helpful when setting these criteria. It would be preferable to clarify the rationale for each criterion. Both Chapter 9 and 10 also demonstrate dilemmas the authors and the students faced when considering inclusion criteria and how they addressed them.

3. Identify search strings for conducting a literature search
   The next stage is to develop a literature search strategy. Search terms should be developed so that they can cover potential articles or documents for the synthesis. Therefore, it may be necessary to include interchangeable terms as well. It is also recommended that a research team be created to agree on the terms. Ideally, the involvement of practitioners would bring an important insight on search strings, making them closer to practice.
4. Conduct a literature search
   Literature search can be conducted across various databases (e.g., Scopus, Web of Science, and ERIC) to find relevant articles. Country-specific databases (e.g., J-Stage and CiNii for topics related to Japan) can also be adopted if necessary. For a practice review, literature can be non-academic (e.g., a government report, a bulletin). Having access to a range of non-academic publications is crucial and thus the encouragement to involve partners in the industry or profession.
5. Screen literature using criteria
   This stage can be divided into two phases: (1) first-level screening and (2) second-level screening. First-level screening is conducted through reading titles and abstracts, based on the inclusion and exclusion criteria. Second-level screening is performed on full texts based on the same set of criteria. At this level of screening, at least two reviewers should be involved in screening a minimum of 10% of the potential studies. The involvement of practitioners as reviewers is recommended, although adequate training needs to be provided. As with Stages 6–8, inter-coder reliability should be at least 80% (Belur et al., 2021). Disagreement among reviewers needs to be discussed and resolved.
6. Appraise study
   Evaluate included studies using a suitable tool or checklist (e.g., Critical Appraisal Skills Programme (CASP). These tools could be adapted for a practice review and qualitative research synthesis. In the case of involving practitioners in the process, it is essential that the tool/checklist is rewritten using jargon-free language.
7. Extract data
   At this stage, one important decision should be made as to whether to extract raw data (i.e., verbatim transcription), interpreted findings (i.e., researchers' interpretations of raw data), or both. This decision is dependent on the type and focus of the synthesis. When the synthesis team encounter missing information and if the missing information is critical for synthesis (e.g., description of an intervention), the team can contact the authors of the study to request additional information. However, if this is not possible, the team should acknowledge this as a limitation or reconsider whether to include the study.

8. Synthesise data
   This is a crucial stage for a qualitative or practice review. Based on the research questions, the extracted data should be analysed qualitatively in a qualitative research synthesis, and qualitatively and/or quantitatively (usually using descriptive statistics) in a practice review. As with the previous stages, practitioners should be involved in the coding and synthesising process to ensure ecological validity.
9. Report synthesised data
   Given the practice-focused nature of the review, reporting of findings is expected to be easy-to-read for practitioners. In addition to creating tables and charts in the article, creating infographics, evidence and gap maps, or other visual aids would be good options to present complex information.

Software/resources

1. NVivo
   This is a software that supports qualitative analysis. It enables its users to code, organise, and retrieve qualitative data. It is especially helpful when conducting a review that requires the handling of a large amount of qualitative data.
2. Rayyan (https://www.rayyan.ai/)
   This is a simple and powerful tool for screening articles. It facilitates collaboration among reviewers when conducting first-level and second-level screening. It also includes AI functions such as deduplication and recommended ratings for inclusion.

Examples of the relevant publications

doi Chen, T. (2016). Technology-supported peer feedback in ESL/EFL writing classes: A research synthesis. *Computer Assisted Language Learning*, *29*(2), 365–397.

This article is an exemplar of a qualitative research synthesis. It provides detailed information on each methodological stage. In the data evaluation stage, in addition to thorough descriptions of how to evaluate literature and code literature, this article specifies how another reviewer was involved in the coding process. The analysis and interpretation stage includes detailed information about how the coding was conducted. The figure and table about data analysis and coding visualise the process and make it easy to understand.

Grima, G., Hooper, A., Mullins, L., Redmond, B., Sharott, P., & Ashworth, B. (2024). *Understanding current practice and research priorities in teaching Writing*. Education Endowment Foundation. https://educationendowmentfoundation.org.uk/education-evidence/evidence-reviews/writing-practice-review

This is a practice review that focuses on current practices in teaching writing. For those who are interested in practice reviews, this publication will be a valuable resource in that it has a thorough description of methodological steps including how the authors handled the extracted data. The data are presented in a reader-friendly way for practitioners. This article summarised information with effective use of tables and figures, which show the findings at a glance.

AI use declaration statement

The author confirms that no AI tools were used in the preparation, writing, formatting, and editing of this work.

References

Belur, J., Tompson, L., Thornton, A., & Simon, M. (2021). Interrater reliability in systematic review methodology: Exploring variation in coder decision-making. *Sociological Methods & Research*, *50*(2), 837–865.

Chen, T. (2016). Technology-supported peer feedback in ESL/EFL writing classes: A research synthesis. *Computer Assisted Language Learning*, *29*(2), 365–397.

Chong, S.W. (2025). Synthesis Methods and Reporting Tool (SMART) for research syntheses in Applied Linguistics. *Research Synthesis in Applied Linguistics*, *1*(1), 17–38.

Chong, S.W., Bond, M., & Chalmers, H. (2024). Opening the methodological black box of research synthesis in language education: Where are we now and where are we heading? *Applied Linguistics Review*, *15*(4), 1557–1568.

Chong, S.W., & Plonsky, L. (2021). A primer on qualitative research synthesis in TESOL. *TESOL Quarterly*, *55*(3), 1024–1034.

CHAPTER 9

Supervising master's students conducting qualitatively-focused research syntheses in a Japanese context

Eoin Jordan[1], Jason Kushnir[2] & Sho Kato[1]
[1] University of St Andrews, Scotland | [2] Independent Researcher, Japan

This chapter reflects on the lead author's experience of supervising two master's students in conducting two separate qualitatively focused research synthesis dissertation projects, both of which were focused on the Japanese educational context. The two supervisees have contributed to this chapter as additional authors, to ensure that it presents a rounded and accurate view of supervisions. Key challenges faced during the dissertation projects are detailed in this chapter; these were overcome iteratively, through ongoing and open discussion between supervisees and supervisor. The chapter should be a useful resource for dissertation supervisors or master's students considering undertaking qualitatively focused research syntheses, particularly those focused on a specific national context.

Keywords: qualitative research synthesis, qualitative synthesis of research, education in Japan, postgraduate supervision, master's dissertation

1. Background

1.1 Introduction

This chapter provides a reflective account of two master's dissertations that conducted Qualitative Research Synthesis (QRS) (Chong & Plonsky, 2021), and its close variant Qualitative Synthesis of Research (QSR) (Liu & Chong, 2024), in a Japanese context. QRS involves the synthesis of only qualitative, or qualitative elements of, research, whereas QSR allows for quantitative research to be synthesised as well, while still maintaining a fundamentally qualitative approach to synthesis. More details on the distinction between QRS and QSR are provided in the section overview for Section 5 of this book. The chapter is written in the voice of, and primarily from the perspective of, the lead author (Eoin), who was the super-

https://doi.org/10.1075/rmal.17.09ajor
Available under the CC BY-NC-ND 4.0 license.© 2026 John Benjamins Publishing Company

visor for these two pieces of research, while also incorporating input from the two researchers/dissertation supervisees themselves, who are the second (Jason) and third (Sho) authors of the chapter. The chapter is intended to provide readers with an overview of the research conducted, as well as insights into some of the challenges faced during the supervised research process and how these were overcome. Additionally, it provides insights into important considerations when conducting QRS or QSR focused on one specific national context, in this case Japan. The chapter should therefore be of particular interest to postgraduate dissertation supervisors or postgraduate students who are thinking about conducting QRS or QSR, as well as practitioners or researchers more generally who are interested in conducting research syntheses in a specific national context.

Jason and Sho, the two researchers whose work is presented here, were distance-learning students on the University of St Andrews MSc TESOL (Teaching English to Speakers of Other Languages) programme when they conducted their dissertation research. I was assigned as supervisor for each of them through two different "matching" processes that paired academic staff with dissertation proposals from students. Both Jason and Sho were based in Japan, and my familiarity with the Japanese educational context, having lived and worked in the country for eight years, was one factor in me being matched as their supervisor. In addition, my interest in educational technologies aligned with the areas of focus of their proposals. Both students had encountered QRS in a research methods module earlier in their master's studies, and indicated an intention to use this method in their dissertation proposals. Jason began his dissertation research in late 2022, making his final submission in April 2023, while Sho began his research in early 2023 and submitted in August 2023.

While I had experience in conducting and supervising qualitative research prior to supervision of the dissertations detailed in this chapter, and was broadly familiar with QRS, I had not previously conducted or supervised research using this specific methodology. Supervision was therefore as much of a learning experience for me as it was for Jason and Sho. I was able to help them with understanding the timeline and expectations for producing a successful dissertation, as well as by providing general guidance on quality in research, and how to report their work appropriately. However, we worked together through information about different forms of research synthesis, as well as occasionally consulting with the first author of Chong and Plonsky (2021), who was a colleague.

Early on in discussing Jason and Sho's research proposals, it was clear to me as their supervisor that their future career paths were likely to be Japan-based, and therefore it would be most meaningful for them to conduct research that would be of particular value in that context. Additionally, my own positioning was (and is) that context is of great importance to any educational research, and

that findings from one research context are not easily transferrable to another one without careful consideration of the similarities and differences between the respective contexts. This belief, broadly aligned with a constructivist research paradigm (Denzin & Lincoln, 2011), is rooted in my own lived experience of being a language teacher in several different countries, teaching students of different ages, in different kinds of institutions. While there have been some aspects of my teaching practice that I have been able to employ successfully in more than one context, other aspects needed to be changed significantly. Working out what was and was not transferrable took careful consideration, reflection, and some degree of trial and error.

My underlying belief in the importance of context, together with Jason and Sho's involvement with the Japanese context, led to discussions in our initial supervision meetings about the possibility of conducting QRS studies based in the national context of Japan, rather than internationally focused syntheses, which had been suggested in their original proposals. Focusing on a single national context also presented as an opportunity to address, to some degree at least, the methodological tension that exists within internationally-focused research syntheses that incorporate qualitative research, such as: Chong and Reinders (2025); Curtis et al. (2025); Whitworth and Rose (2025); Zingaretti and Frear (2025). The tension is that these studies, while providing findings that are interesting and useful to both practitioners and researchers, are taking qualitative research studies typically based in a constructivist paradigm, which assumes findings cannot be generalised to other contexts, and seeking to synthesise these findings across multiple different contexts. While there is still variation within any national context, narrowing the focus of a research synthesis to a single national context appeared to be one way of reducing the methodological tension noted here, while also yielding findings that would be of strong relevance to practitioners and researchers working in that context. This issue is explored in more detail in Sections 2.1.3, 3.3 and 4.3 in this chapter. Chapter 10 of this book also explores the challenges of employing a slightly wider geographical/cultural range (the Anglosphere) to a research synthesis project.

The remainder of this introduction provides an overview of Jason's and Sho's dissertation research projects, presented in chronological order. Next, key methodological and ethical considerations encountered before and during the process of synthesis are detailed, and following that, an account of how these considerations were addressed is shared, as well as alternative possible courses of action. The chapter ends with a summary of lessons learned, some questions for further contemplation, and some recommended further reading.

1.2 Overview of Jason's dissertation research

Jason worked as an English teacher at multiple institutions in Japan at the time he was working on his dissertation. His dissertation research focused on barriers to education technology uptake in post-secondary education in Japan, with the title "Barriers to educational technology in Japanese education: a qualitative research synthesis of post-secondary teacher views". This area of focus was rooted in some of his own experiences working as a teacher in Japan, and noting variable uptake of educational technology, including through the COVID-19 pandemic, when greater use of this technology was necessitated. QRS was adopted as an approach because of the limited number of articles that were expected to meet the inclusion criteria for the research, and also because it was manageable within the time period allowable for a dissertation. His research examined teacher perceptions reported in 17 qualitative and mixed methods studies. Results generated from a thematic analysis (Braun & Clarke, 2006) of the primary data suggested that teacher confidence in the implementation of educational technology was an important barrier, which may be exacerbated by a perceived lack of support and training.

After successfully completing his dissertation, Jason published this on the OSF preprints site (Kushnir, 2023). To date (4 December 2025), it has received 230 views and 365 downloads.

1.3 Overview of Sho's dissertation research

At the time of his dissertation, Sho was working as an English teacher in a high school in Japan. He was interested in the use of technology to support vocabulary learning, noting the pressure on high school students to learn greater volumes of English vocabulary under recent English curriculum reforms introduced by the Japanese Ministry of Education, Culture, Sports, Science and Technology (abbreviated to MEXT) in 2018 (MEXT, 2018), and therefore the pressure on teachers to find ways to facilitate this. His dissertation, titled "Technology-mediated Vocabulary Learning in EFL Classrooms in Japan: A Qualitative Synthesis of Research" ultimately ended up being a qualitative synthesis of research, rather than a qualitative research synthesis — the distinction and rationale for this change is explained later in this chapter (see 2.2.1 and 3.6). As with Jason, Sho had initially intended to conduct a QRS, based on expectations regarding the number of articles that would meet the inclusion criteria, and the limited time period available to complete his dissertation research. Data was extracted from 17 studies, and then synthesised using thematic analysis (Braun & Clarke, 2006). The

results highlighted characteristics of technology-mediated vocabulary learning in a Japanese context, as well as affordances and limitations.

After successfully completing his dissertation, Sho also published this on the OSF preprints site (Kato, 2023). To date (4 December 2025), it has received 206 views and 189 downloads.

2. Methodological and ethical considerations

This section will outline some key methodological and ethical considerations that were relevant to Jason's and Sho's research. These are separated into items that were considered prior to the main research process beginning, which accounts for the majority, and then one issue that arose during the research process itself.

2.1 Considerations before the research began

2.1.1 *Practicality and ethics in an age of data abundance*

One practical issue that attracted both Jason and Sho to consider research synthesis for their dissertations was that it would not require collecting new data from human participants, and the ethical approval and informed consent processes associated with this. In particular for Sho, who was working at high school level with many students who are under the age of 18, obtaining informed consent would have presented additional challenges relative to collecting data from adults, as collecting data from this group of non-adults may have required parental consent. For both Sho and Jason, the short duration of the dissertation research period (approximately three months), and the need to formulate a research plan and then gain ethical approval for this before commencing data collection, meant that the actual window for collecting data would have been small. Data collection would need to have finished far enough ahead of the final submission deadline to allow sufficient time for analysis and writing up, further compressing the potential data collection period. While ultimately navigating ethics approval processes and obtaining informed consent from participants would likely have been achievable for both Sho and Jason, it would have taken time away from analysis, relative to a research synthesis project where there were fewer ethical "hurdles" to clear.

Practicality was also a central consideration in the decision to undertake a qualitative form of research synthesis, rather than a more quantitatively focused approach such as meta-analysis. As noted in 1.1, Jason and Sho had encountered QRS in one of their research methods modules during their master's study, and already had a foundational understanding of this method, and its potential to

provide useful information for practitioners. While a more quantitatively focused method could have been considered, this would have involved a steeper learning curve for both researchers. If they had conducted a meta-analysis then they would have needed to do a considerable amount of additional study on statistical techniques, over and above the learning they had done in this area on their research methods modules, to enable them to engage with this method effectively and robustly.

Additionally, from my perspective as a supervisor, an important ethical consideration was that we are now living in a world of data abundance. Indeed, this has been noted by other authors as a reason for the importance of research synthesis (Gusenbauer, 2021). Given the large amount of published data on many topics that already exists, and the nontrivial investment of time required by both researchers and participants to generate new data, increasing our overall research focus on synthesising and understanding data that already exists seems both an ethical and useful approach. That is not to say that there is no value in generating any new data, but rather that the collection of new data from human participants should be done selectively, with a strong focus on quality and on avoiding 'research waste' (Glasziou & Chalmers, 2018), after considering the possibility that a meaningful contribution could also be made to the field of interest by synthesising published data.

2.1.2 *Supervisory ethics*

A further ethical consideration from my perspective as supervisor was my responsibility to guide and support Jason and Sho in conducting research that had relevance to their interests, but also to their current and future working lives. Both researchers had busy jobs, as well as other personal commitments outside of these, so it was important from my perspective to encourage them to pursue research topics that were meaningful to and in some way synergetic with their work, and that would also potentially help them to develop expertise to support future career development. Ultimately, they had enrolled on the MSc TESOL programme in large part to support their career development, so as their dissertation supervisor, I felt a strong sense of responsibility to ensure that we did not lose sight of this broader goal when considering the topic and approach for their research.

2.1.3 *Methodological tension between qualitative research and research synthesis*

As noted in 1.1, one methodological consideration that arose early on in my discussions with Jason and Sho about their plans to conduct a research synthesis was the width of the contextual focus within this. The research syntheses both

Jason and Sho had encountered through their research methods module study experiences were internationally focused; however, in our initial meetings, we discussed the nature of qualitative research, and the typical situation of this type of research within a constructivist paradigm (Denzin & Lincoln, 2011), where findings are usually not intended or believed to be generalisable across contexts. I suggested that one way of addressing, to some degree at least, this methodological tension could be to focus their research syntheses on one specific context, to reduce the contextual differences between the studies being synthesised. Our discussions considered that one important dimension for this contextual "narrowing" could be geography, and the idea emerged of conducting a research synthesis focused either on a geographical region that may share some cultural similarities, such as Confucian Heritage Countries (Evers et al., 2011), or on a specific country, such as Japan.

2.1.4 *Supervisee and supervisor roles within the research*

One further consideration that had both methodological and ethical dimensions was the degree to which I could, as a dissertation supervisor, collaborate with Sho and Jason in their research. Typically, research syntheses are conducted by more than one person, and this is advised in recent guidelines (Chong, 2025); however, dissertation supervisors within the International Education and Lifelong Learning Institute (IELLI) at University of St Andrews need to adhere to guidelines that limit the degree to which feedback and support is given to supervisees, to ensure consistency and fairness in how the whole cohort of students is treated. As the dissertation is an individual piece of work, which is ultimately graded by the supervisor together with a second marker, it would not be appropriate or ethical for either a supervisor or anyone else to have a large input into the projects. This tension between me playing a significant part in the research process to ensure the robustness of that process, and the need to ensure consistency in the master's dissertation assessment process as well as confidence that the research could reasonably be considered the individual work of the student, was an important consideration when I discussed with both Jason and Sho the roles that we each would play at each stage of the research process.

2.1.5 *Publication languages to be used in the synthesis*

Another consideration that arose and was discussed with both Jason and Sho prior to the research process commencing was the publication languages that would be included in their respective syntheses. Examples of QRS that they had seen employed inclusion criteria that limited included sources to those that were published in English (e.g. Chong & Reinders 2020). Given the dominance of English as a language for international peer-reviewed journals, this made sense for

an internationally-focused QRS, but once we began discussing the possibility of research syntheses focused specifically on Japan, this prompted a question about whether or not publications in the "local" language of Japanese should also be included. A substantial amount of academic literature related to language teaching in Japan is published in Japanese, such as articles in in-house Japanese university journals, and this body of literature could shed additional light on the field of interest within Jason and Sho's research.

When discussing what languages of publication should be included in the respective research syntheses, practicality was a consideration as well. The researchers would need to be confident in engaging in detail with Japanese language academic literature if they were to incorporate this into their studies. If included, consideration was then also needed regarding whether this should have any impact on the language employed when coding texts.

2.2 Considerations that arose during the process of synthesis

2.2.1 *Initial inclusion criteria not yielding sufficient results*

While the majority of the potential issues that could arise were discussed by the researchers and me prior to beginning the research process, one issue that emerged after this process had begun was that searches using the initial inclusion criteria specified by both Jason and Sho did not yield sufficient numbers of publications to conduct a meaningful research synthesis. This only became clear after each researcher had reviewed titles and abstracts of publications, and had noted that while their initial search had yielded a significant number of results, on closer examination based on the inclusion criteria, only a relatively small number of these met all the inclusion criteria.

Following this outcome, I discussed with each researcher what elements of the inclusion criteria they could consider adjusting to allow for a larger number of texts to be included, while also preserving the integrity of the research.

3. Course of action

This section will detail how the methodological and ethical considerations noted in Section 2 were addressed.

3.1 Practicality and ethics in an age of data abundance

As indicated in 2.1.1, Jason and Sho both already had an intention to conduct a QRS study at the point when I became their supervisor. However, our discussions about the practicality of this option for a student working on a time-limited dissertation, as well as the broader ethical point (also covered in 2.1.1) relating to the abundance of data in existence, helped to affirm this intention. After careful consideration, both researchers decided to move ahead with QRS (although in Sho's case, this later switched to QSR).

3.2 Supervisory ethics

In the early stages of supervision, I made sure that I was responsive to the intentions and interests of both researchers, while also raising and discussing with them how their planned research projects might be useful to their own educational contexts and future career development. For Jason, his intended focus on barriers to educational technology was rooted in his lived professional experiences during the COVID-19 pandemic, but also reflected an ongoing interest in educational technology, and an awareness of Japanese governmental moves to introduce greater technology into society. With this in mind, it seemed likely that this focus would be relevant to his future professional activities in education in Japan.

Sho viewed technology as a tool that may be useful to aid students with English vocabulary learning, which he had identified as a significant challenge for students in his particular context, as a result of MEXT changes to English language education and elementary, junior high and high school levels. His research plans were developed with an awareness that challenges around English language vocabulary learning are likely to be an important concern within the Japanese educational system moving forward, and therefore that his dissertation research would be of relevance to his future professional activities.

3.3 Methodological tension between qualitative research and research synthesis

Following my discussions with Jason and Sho regarding the methodological tension between QRS and the nature of much qualitative research (described in 2.1.3), as well as the possibility of addressing this by narrowing the geographical context of their research syntheses, both researchers ultimately decided to limit the scope of their work to Japan. While the broader region of East Asia was discussed as a possibility too, and one which would likely have yielded a larger body

of literature for potential inclusion, the researchers and I felt that there was significantly more coherence in just focusing on Japan. Limiting the scope to one country limited cultural variations and variations in the educational system within the research, which would have been more diverse if the literature had been synthesised across multiple countries, even if those countries shared some cultural heritage. Since supervising Jason and Sho, I have also taken a similar approach in another collaborative QSR project focused on the Japanese context (Stringer & Jordan, 2025), for similar reasons. For both Jason and Sho, the focus on Japan was relevant to their future careers too — a consideration discussed in 2.1.2 and 3.2. The challenges of synthesising literature from a wider realm (the Anglosphere) are explored in more detail in Chapter 10 of this book.

3.4 Supervisee and supervisor roles within the research

After careful discussion about finding a balance between preserving the integrity of both the research process and the dissertation assessment process, we agreed that an appropriate approach would be for me as supervisor to act as a "second pair of eyes" on the data extraction and coding that Sho and Jason undertook. We agreed that, while the researchers would undertake all data extraction and coding themselves, I would review a sample of the data extraction and provide feedback on this for them to act on. At a later stage in the process, I would then review a sample of the initial codes developed by the researchers, with reference to the original texts that were coded, and provide feedback on this too. Both reviews noted here were undertaken within the standard framework of holding six hour-long supervision meetings, and communication in between these meetings via a private Microsoft Teams channel, to ensure consistency in approach with other dissertation students. As supervisor, I offered advice and feedback on other parts of the research process too (e.g. development of research questions and inclusion criteria), but it was the data extraction and coding stages where my review was particularly in depth.

Jason and Sho recorded our discussions about data extraction and coding (as well as other aspects of the research process) in meeting notes. Together with Microsoft Teams asynchronous discussions, these became a record of evidence to show that a rigorous and carefully considered research process had been implemented. For example, after our third meeting, Jason wrote the following in his notes about article selection:

> Selected articles — Eoin suggested that we go through the complete if tentative set of selected articles ($n = 20$), based on inclusion criteria aligned with the QRS. In particular:
>
> Participants. It was noted that in three articles the participants were:
> a. not necessarily involved in EFL (CFL, JFL)
> b. in vocational schools rather than HE
>
> One additional article was flagged because the EFL focus was not clearly established. (The EFL focus was however confirmed after the meeting.)
>
> (Jason, notes from third dissertation supervision meeting)

As a further example, after our fifth meeting, Sho wrote the following in his meeting notes about themes identified through the coding process:

> There are some similar themes and those can be put together (e.g., difficult and complicated, helpful and favorable, effective and efficient, etc.).
>
> (Sho, notes from fifth dissertation supervision meeting)

3.5 Publication languages to be used in the synthesis

Careful consideration was given to the languages of publication that would be permissible within the inclusion criteria for Jason and Sho's research. Both syntheses were focused on English language teaching, so given this, together with the dominance of English globally as a language of academic publishing, it made sense in both cases to include publications written in English. Given the Japan-based nature of their research, both researchers recognised the value of including Japanese publications within their searches; however, as the dissertation is an individual piece of research, this also necessitated the researchers needing to be confident that they could engage in depth with Japanese language literature by themselves.

In Jason's case, despite having a good general level of Japanese language ability, he did not feel confident that he would be able to do this independently, and he was ultimately able to find sufficient literature in English to support his research. Sho was confident in his ability to engage with Japanese literature, and so included this within his literature search. For Sho's bilingual research, we also then considered how this might affect the languages used at different stages of the research process. We agreed that Japanese text should be quoted directly in data extraction forms, to preserve the original text at that stage. However, we also agreed that coding on Japanese texts should be done in English, to allow clearer comparability with coding on other publications within the sample.

3.6 Initial inclusion criteria not yielding sufficient results

Both Jason and Sho needed to adapt the inclusion criteria for their studies after finding that their initial literature search did not yield enough results to proceed with a meaningful synthesis. In Jason's case, he had initially limited the scope of his QRS to higher education, but after further discussion, he opted to widen the inclusion criteria to allow studies at other levels of education — this resulted in the inclusion of one publication from a post-secondary vocational setting.

For Sho, he found that there was a lack of qualitative, or primarily qualitative, research that met his initial inclusion criteria. He then considered expanding the scope of his criteria in different dimensions. After discussion with me, he chose to change his research from being a QRS to a QSR. This allowed for the inclusion of findings from quantitative research, which would then be qualitatively synthesised. Ultimately, both Jason and Sho selected different inclusion criteria to "loosen" requirements, based on an evaluation of the similarity of contexts (in Jason's case), and following the approach taken by other researchers who faced a similar issue in Sho's case (Liu & Chong, 2024).

4. Alternative courses of action

This section will explore what other courses of action could have been available to address the considerations noted in Section 2, as well as (where appropriate) what the consequences may have been for the research and researchers if these courses of action had been followed.

4.1 Practicality and ethics in an age of data abundance

While Jason and Sho were already quite committed to QRS at the start of supervision, we did spend some time initially discussing alternative possibilities for their dissertation research. In particular, in line with their interest in qualitative rather than quantitative research, we discussed the possibility of conducting interview studies to explore the topics that were of interest to them. However, for reasons already presented in 2.1.1 and 3.1, this approach was not pursued.

4.2 Supervisory ethics

It would have been possible for me in my supervisory role to have encouraged supervisees to pursue a research topic that interested them, without giving consideration to relevance to their future careers. Alternatively, I could have encouraged

them to consider cognate areas of research to their initial proposals that aligned more closely with my own research interests — for example, vocabulary assessment for Sho, or online peer assessment for Jason. Doing the second of these would arguably have allowed me to provide them with greater insight into their research topic, but ultimately I did not feel this approach would have been to the benefit of the researchers and their future careers.

4.3 Methodological tension between qualitative research and research synthesis

Jason and Sho could both have considered conducting an internationally-focused research synthesis, rather than limiting the scope of their work to Japan. Indeed, Sho's initial proposal was to do this; however, discussion of methodological tension between qualitative research and synthesising this across a wide range of contexts, the consideration that his future career was likely to be Japan-based, and the prospect of finding a clearer niche for his research resulted in him limiting his focus to Japan. Jason had tentatively suggested in his initial proposal that the synthesis he would do would be primarily focused on East Asia or Japan, and after initial discussions, he narrowed this down to Japan. This also was in recognition of him likely to continue to work in Japan, as well as the point about methodological tension, especially when considering the differences between educational systems across East Asia. Had the researchers pursued internationally-focused research syntheses, this would likely have increased the number of publications that met their inclusion criteria, therefore making that part of the research process easier to navigate. However, from my perspective, it would also have resulted in less cohesive and meaningful (particularly to the researchers themselves) pieces of research than what they actually produced.

4.4 Supervisee and supervisor roles within the research

When considering my involvement with the data extraction and coding stages of the research process, two other possibilities would have been for me to do some of the data extraction and coding, or for me to recommend that the researchers found other collaborators to do this. I felt that the former of these would not have been an appropriate or ethical approach, given expectations at St Andrews for the research master's students to conduct their dissertations entirely on their own, and that the supervisor is the first marker for the dissertation. It is possible that this approach could be viable in contexts where guidance allows for the supervisor to play a greater role in the research process, and where the supervisor is not involved in marking dissertations for the students they supervise; however, the

time commitment required of the supervisor to engage in detail with data extraction and coding would also need to be considered.

The second suggestion would avoid the potential conflict of interest in the supervisor ultimately being a marker for the work; however, it presented other potential challenges. If this approach was adopted, there could potentially be a question around the degree to which the dissertation research had been entirely conducted by the student themselves. Additionally, it would have been challenging to source someone with appropriate expertise who would have the time and necessary commitment to conduct data extraction and coding within the tight timeframe required by the dissertation period. In my particular context, other master's students or doctoral students could be possible candidates, but these are typically people who are already busy with their own studies, and in the case of distance-learning students, often full-time employment and family commitments as well. For these reasons, I did not guide Jason or Sho towards this course of action.

4.5 Publication languages to be used in the synthesis

As noted in 3.5, Jason was not sufficiently confident in his ability to engage with Japanese language literature to include this within his synthesis. The usability and accuracy of translation software have increased considerably in recent years (Jolley, 2022), so one possibility would have been to employ this to assist him in engaging with Japanese language publications. However, despite recent advances, the use of translation software still seems likely to increase the probability of misinterpretation of text content. This trade-off between widening the range of literature that could be reviewed and the depth and accuracy with which this literature could be analysed needs to be given careful consideration in research designs such as this.

4.6 Initial inclusion criteria not yielding sufficient results

For both Jason and Sho, there were alternative possible responses to the lack of publications that met their initial inclusion criteria. Firstly, they could have chosen other inclusion criteria to loosen instead of the ones they did. For instance, although it would conflict with the aim of reducing the methodological tension inherent in internationally-focused QRS, they could have broadened the geographical range of settings with the criteria. One idea that was discussed with both researchers was extending the geographical range to East Asia, or to Confucian Heritage Countries (Evers et al., 2011), to emphasise a group of countries or territories with some degree of similar cultural heritage. These changes would potentially have enabled the inclusion of significant bodies of literature from China, Hong Kong, Taiwan and South Korea.

However, taking this approach would have weakened the intended approach of focusing on a narrow context, would have taken focus away from Japan within the research (which was the site of both researchers' main experience working in education, and also likely to be the site of their future careers), and in Sho's case it would also have meant that he would not be able to adopt a bilingual approach to literature search in locations other than Japan. This would have generated a methodological inconsistency, or would have required him to adopt an English-only search strategy across all countries/territories, which would have further limited the research he could access and include from the Japanese context.

One further option available to the researchers was to consider changing to a different type of research synthesis. For example, a meta-analysis could have been one option for Sho, having found that there was a strong presence of quantitative research with effect size calculation on his topic area. However, there were obstacles to this approach in that neither Jason nor Sho had been prepared to conduct a meta-analysis through their research methods modules in their master's programme. The learning curve to conduct a rigorous meta-analysis would have been quite steep for both the researchers, and for me as a supervisor. Beyond this, through discussions with Jason and Sho, it was clear that they were committed to taking a qualitative approach to research, in view of the accessibility and meaningfulness of this for practitioners.

5. Conclusion

5.1 Summary

This chapter has provided a brief overview of the QRS conducted by Jason and the QSR conducted by Sho as their master's dissertation research projects. It has then delved into the experience of supervising, and being supervised, when conducting this type of research for a master's dissertation. It has explored some important methodological and ethical considerations that arose both prior to and during the main QRS/QSR research process. These were: considerations of practicality and ethics regarding the initial choice of research approach; supervisory ethics regarding the research focus and scope; methodological tension between qualitative research and research synthesis; supervisee and supervisor roles during the research process; publication languages to be included in syntheses; and adjustments to inclusion criteria for syntheses.

5.2 Lessons learned

Reflecting on the experience of supervising the two dissertations detailed in this chapter, there are several key lessons that I (Eoin) have learned about how to approach supervision of QRS (or QSR). These align broadly with items discussed up to this point in the chapter, and I hope they will provide a useful guide for both supervisors and supervisees embarking on similar research projects.

Firstly, for any dissertation research project (including research synthesis), it is important to consider the practicalities of conducting the research early on, taking required timelines into account. It is also important to give explicit consideration to ethical issues that lie behind decisions to adopt a particular research design, even if this does not involve collecting data from human participants.

Secondly, in a supervisor role it is important to be guided primarily by what is in the best interests of supervisees, taking into account both their present situation and their planned future activities, rather than guiding supervisees too strongly towards following one's own specific research interests.

Thirdly, consideration should be given to the philosophical underpinnings of research to be synthesised when undertaking research synthesis. In the case of QRS, this does not mean taking a hard-line stance that it is not meaningful to synthesise qualitative research which has been conducted in disparate contexts, but instead to explicitly recognise the methodological tension within the method, and take pragmatic steps to minimise contextual differences between the studies being synthesised where possible.

Fourthly, supervisor and supervisee roles within the research process need to be agreed in detail early on during supervision. Care needs to be taken to strike a balance between ensuring alignment with institutional policy surrounding dissertation/research supervision, while also ensuring that a robust research process is implemented.

Fifthly, the languages of publication that will be included in a QRS project need to be considered carefully before the research process begins. In the case of nationally-focused research syntheses, such as those described in this chapter, there may be a compelling argument to include publications in both English and a local language (where this is not English). However, researchers' language abilities also need to inform decision-making in this area, and where texts in multiple languages are synthesised, there needs to be a clear plan for which language(s) will be used at each stage of the research process (in particular, data extraction and coding).

Finally, it is advantageous to make contingency plans for the scenario that initial inclusion criteria do not yield sufficient results to conduct a meaningful research synthesis. A sensible approach might be to label each of the inclusion criteria in terms of priority, with a view to then adjusting the lowest priority criterion first if enough publications are not found.

6. Questions for discussion

This section poses some useful questions that supervisors and supervisees may wish to reflect on and discuss if they are considering a QRS/QSR project, or a research synthesis project more generally.

- **Why synthesise qualitative research?** Researchers can reflect on whether their motivation is rooted in a worldview that values qualitative over quantitative research, a desire to provide practitioners with findings they can apply in their professional context, familiarity with the qualitative research synthesis process, or some other reason or combination of reasons.
- **Is there methodological consistency between the type of synthesis being conducted and the relative wideness or narrowness of the educational contexts being considered?** As discussed in this chapter, there is a potential methodological tension between a focus on qualitative research, which does not normally aim to generate results that are generalisable to other contexts, and the synthesis of findings from qualitative research conducted in different contexts. It is important for supervisees and the supervisors to consider how they can reduce the "distance" between the contexts of research being synthesised through their inclusion criteria, while also weighing up this consideration against other factors (e.g. measures of research quality).
- **What roles should supervisor and supervisee play within the research synthesis process, to ensure robustness, but also to ensure that the research has been carried out in line with institutional policies for postgraduate supervision?** Consideration needs to be given in the early stages of supervision in particular to the role that supervisors will play at the data extraction and coding stages of the research synthesis process.
- **What languages will be part of the synthesis?** While many syntheses focus exclusively on English language publications in international peer-reviewed journals, this excludes literature published in local languages, which may provide deeper insights into local practices. In cases where supervisees possess linguistic ability in local languages, consideration should be given to whether this ability can be used to contribute to the research.
- **What are the most and least important items within your inclusion criteria?** Supervisees need to be prepared to make changes to the inclusion criteria after initial exploratory searches, in the case that there is not a sufficient number of publications that meet these criteria. It is helpful to have considered in advance which of the inclusion criteria should be the first to be adjusted in this situation.

Further reading

Chong, S.W., & Plonsky, L. (2021). A primer on qualitative research synthesis in TESOL. *Tesol Quarterly*, *55*(3), 1024–1034.

This article provides an overview of QRS within the field of TESOL, as well as a clear and easy-to-follow process and example for conducting this type of synthesis.

Liu, Q., & Chong, S.W. (2024). Bilingual education in China: A qualitative synthesis of research on models and perceptions. *Applied Linguistics Review*, *15*(4), 1671–1693.

This article is a useful exemplar of QSR, adopting a methodology adapted from Chong and Plonsky (2021).

Stringer, T., & Jordan, E. (2025). A qualitative synthesis of practitioner research of English language learner autonomy in higher education in Japan. *Research Synthesis in Applied Linguistics*, *1*(2), 414–458.

This article is a useful exemplar of TESOL-related QSR being conducted in a single national context (Japan).

Chong, S.W. (2025). Synthesis Methods and Reporting Tool (SMART) for research syntheses in Applied Linguistics. *Research Synthesis in Applied Linguistics*, *1*(1), 17–38.

This article offers a helpful guide on methodological and reporting practices for all forms of research synthesis in the field of Applied Linguistics.

AI use declaration statement

The authors confirm that no AI tools were used in the preparation, writing, formatting, and editing of this work.

References

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101.

Chong, S.W. (2025). Synthesis Methods and Reporting Tool (SMART) for research syntheses in Applied Linguistics. *Research Synthesis in Applied Linguistics*, *1*(1), 17–38.

Chong, S.W., & Plonsky, L. (2021). A primer on qualitative research synthesis in TESOL. *Tesol Quarterly*, *55*(3), 1024–1034.

Chong, S.W., & Reinders, H. (2020). Technology-mediated task-based language teaching: A qualitative research synthesis. *Language Learning and Technology*, *24*(3), 70.

Chong, S.W., & Reinders, H. (2025). Autonomy of English language learners: A scoping review of research and practice. *Language Teaching Research*, *29*(2), 607–632.

Curtis, K., Chong, S.W., & Kong, M.S. (2025). A qualitative synthesis of research into the use of exemplars in the English for academic purposes context to develop student feedback literacy. *Research Synthesis in Applied Linguistics*, *1*(2), 302–339.

Denzin, N. K., & Lincoln, Y. S. (2011). *The Sage handbook of qualitative research*. Sage.

Evers, C. W., King, M., & Katyal, K. R. (2011). Conducting research in Confucian Heritage Cultures: An overview of methodological issues. *Comparative Education*, *47*(3), 295–300.

Glasziou, P., & Chalmers, I. (2018). Research waste is still a scandal — An essay by Paul Glasziou and Iain Chalmers. *BMJ*, *363*, k4645.

Gusenbauer, M. (2021). The age of abundant scholarly information and its synthesis– A time when 'just google it' is no longer enough. *Research Synthesis Methods*, *12*(6), 684–691.

Jolley, J. R., Maimone, L. (2022). Thirty years of machine translation in language teaching and learning: A review of the literature. *L2 Journal*, *14*.

Kato, S. (2023). Technology-mediated vocabulary learning in EFL classrooms in Japan: A qualitative synthesis of research (Unpublished master's dissertation). University of St. Andrews.

Kushnir, J. (2023). Barriers to educational technology in Japanese education: A qualitative research synthesis of post-secondary teacher views (Unpublished master's dissertation).

Liu, Q., & Chong, S. W. (2024). Bilingual education in China: A qualitative synthesis of research on models and perceptions. *Applied Linguistics Review*, *15*(4), 1671–1693.

MEXT. (2018). Koutou gakkou gakusyuu sidou youryo, gaikokugo hen, eigo hen [*The Course of Study for high school. Part of English as foreign language*]. https://www.mext.go.jp/content/1407073_09_1_2.pdf

Stringer, T., & Jordan, E. (2025). A qualitative synthesis of practitioner research of English language learner autonomy in higher education in Japan. *Research Synthesis in Applied Linguistics*, *1*(2), 414–458.

Whitworth, B., & Rose, H. (2025). A systematic review of research on the use of shadowing for second language pronunciation teaching. *Research Synthesis in Applied Linguistics*, *1*(2), 239–269.

Zingaretti, M., & Frear, L. (2025). The benefits and challenges of queering TESOL: A qualitative research synthesis. *Innovation in Language Learning and Teaching*, 1–23.

CHAPTER 10

Conducting a research synthesis on translanguaging practices in Anglophone classrooms

Ming Sum Kong[1,2] & Sarah Hopkyns[1,3]
[1] University of St Andrews | [2] University of East London | [3] University College London

This chapter discusses the process and experience of conducting a research synthesis on translanguaging practices in Anglophone classrooms. Translanguaging challenges monolingual policies in many educational contexts and encourages language learners to utilise their full linguistic repertoires to learn. Although the term 'translanguaging' is widely discussed, its conceptualisation and applications to practice are still varied, which can make translanguaging activities in the classroom challenging. Our research synthesis takes the form of a 'practice review', where we assess relevant literature in both academic journals and practitioner magazines. The review aims to identify and gain insights into 'what translanguaging looks like in practice' and 'how it can be applied as a pedagogical approach in real-life classrooms'. The chapter starts by introducing the rationale and aims of the practice review. We then reflect on the dilemmas we faced during the research process such as limited results found in the literature and defining the Anglosphere. Based on these dilemmas, we discuss what we did and what we could have done differently. The chapter ends with discussion questions and a list of recommended reading. Through this reflection process, we discuss both challenges and benefits of conducting a practice review, with examples from our own experiences.

Keywords: translanguaging, practice review, Anglophone, collaborative reflection, research synthesis

1. Background

This ongoing synthesis project is a practice review aiming to bridge the gap between researchers' and practitioners' knowledge on translanguaging pedagogical practices through reviewing and examining the existing literature (see Kong

https://doi.org/10.1075/rmal.17.10kon
Available under the CC BY-NC-ND 4.0 license.© 2026 John Benjamins Publishing Company

et al., 2024, for more details). Given the nature of a practice review, we not only included academic articles, but we also included articles from professional magazines, such as the 'International Association of Teaching English as a Foreign Language (IATEFL) Voice', 'English Teaching Professional', and 'Modern Language Teacher', in our analysis. Through this analysis, we aimed to classify translanguaging practices and discover what translanguaging practices look like in Anglophone classrooms. In this chapter, we will discuss the rationale of the project, dilemmas that we encountered and our courses of action. We will end this chapter with questions for discussion and some suggested readings.

Translanguaging can be spontaneous or planned as a pedagogical approach which encourages learners to use their full linguistic repertoires, including semiotic resources, flexibly for communication and learning, especially multilingual learners (Vogel & García, 2017). As mentioned in Chapter 1 of this volume, the 'multilingual turns' has recently gained a lot of attention among scholars and translanguaging is one of the practices that reflects this turn. This approach has been widely discussed and studied in different educational contexts. Nonetheless, its conceptualisation and applications to practice are varied, which can make planned classroom activities using translanguaging difficult to categorise, and, therefore challenging. Added to this challenge is the fact that the term translanguaging is often compared with or used interchangeably with related terms such as code-switching, plurilingualism, translingual practices, and polylanguaging.

Bonacina-Pugh et al. (2021) conceptualised translanguaging by outlining two main approaches: the fixed language approach and the fluid language approach. They created a conceptual framework (see Figure 1) based on García and Lin's (2017) concept of strong and weak forms of translanguaging. The fixed language approach refers to planned language use either by teachers or students in classrooms, while the fluid language approach describes the unplanned, spontaneous use of language in the setting (Bonacina-Pugh et al., 2021). The latter challenges monolingual practice in classrooms and the boundaries of named languages. It argues that bilinguals strategically select appropriate features from one repertoire to communicate effectively in specific contexts (García, 2012; García & Lin, 2017). Bilinguals' full repertoire is not limited to named languages or verbal communication. It includes semiotic, non-verbal communication and visual aids (Bonacina-Pugh et al., 2021; García & Otheguy, 2020). In fact, translanguaging has been utilised and studied in sign language contexts too (Kusters et al., 2017). However, even though Bonacian-Pugh et al.'s (2021) framework further deconstructed the concepts and practices of translanguaging, it does not inform us of how these approaches can be applied in real-life classrooms, such as what instruction and tasks can be given to students and what languages or forms of communication would be involved. It, therefore, arguably provides limited practical

insights for practitioners. Treffers-Daller (2025) points out that, apart from the classification García and Lin offered in 2017, further classifications of translanguaging are needed. She also argues that due to the ambiguous conceptualisation and definitions of translanguaging, such as repertoires and translanguaging spaces, translanguaging practices in the field can be varied. This makes it difficult for researchers and practitioners to investigate and apply pedagogical translanguaging practice. Therefore, creating 'clear diagnostic criteria' (Treffers-Daller, 2025, p. 4) for translanguaging is important and reviewing and examining current translanguaging classroom practices would be useful.

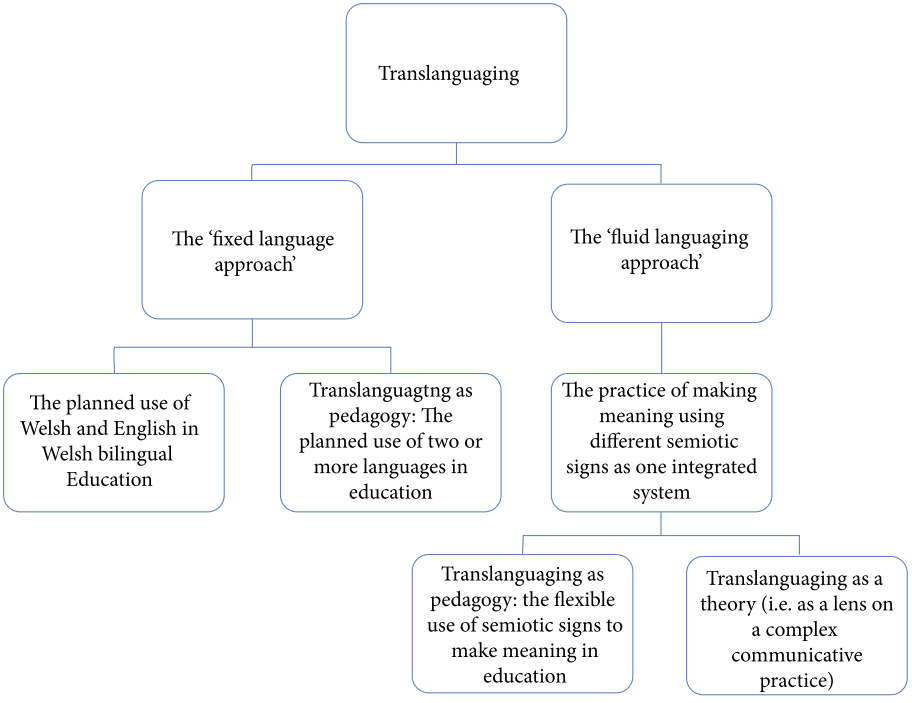


Figure 1. The conceptual map of 'translanguaging' (Bonacina-Pugh et al., 2021, p. 450)

Although language use in Anglophone Higher Education (HE) contexts can be complex, which makes them a rich resource for investigation due to their high international populations, most of the studies on translanguaging have been conducted in non-Anglophone countries. Prilutskaya's (2021) review shows nearly 60% of the related studies were based in non-English speaking countries, for instance, in Europe, Africa and Asia; the only three Anglophone areas were North America, Australia and New Zealand. This suggests that translanguaging practice seems to be under-researched in the Anglosphere. In fact, given that translanguag-

ing challenges monolingual practice and suggests the flexible use of languages in classrooms, it has been a popular topic within English Medium of Instruction (EMI) settings in the past decade. In an EMI context, English is used as an instructional language between students who have English as their second language, while the language itself is not the subject being taught nor intentionally developed (Pecorari & Malmström, 2018). However, based on this definition of EMI, countries with English as their first language are excluded in this context. Research on translanguaging in EMI is often from countries that are in the Outer and Expanding Circles of Kachru's (1985) three concentric circles of World Englishes, where English is the second language and/or a foreign language in the country, such as in China (Wang & Curdt-Christiansen, 2019), Portugal (Caruso, 2018), Spain (Llanes & Cots, 2022), Taiwan (Chang, 2019), and the United Arab Emirates (Hopkyns, 2022). It is worth noting that the distinction between Kachru's concentric circles is blurring in today's era of globalisation and increased movement within and between nations. Indeed, Anglophone contexts in the so-labelled 'inner circle' are linguistically diverse, especially with regard to internationalised universities. Thus, there is a high potential for researching translanguaging pedagogical practice within the Anglophone context. It is also possible that research done in these contexts has been overlooked. Dafouz and Smit (2020) argue that 'English Medium Education in Multilingual University Settings (EMEMUS)' is a more inclusive term as it expands the scope of EMI to capture the dynamic and diverse multilingual environments in the current HE settings. This further supports the point that contexts outside traditionally defined EMI contexts and their research could be overlooked due to the imbalanced attention paid to EMI contexts with the traditional definition. Therefore, in this review, we focus on translanguaging practice research within Anglophone countries.

In addition to showing that research on translanguaging in English-speaking countries has been limited, Prilutskaya (2021)'s systematic review also demonstrates that only 25% of the studies were conducted in HE settings. It is possible that studies from Anglophone HE settings were excluded in the review due to their failure to meet the inclusion criteria. Nevertheless, Prilutskaya's (2021) results are further supported by Huang and Chalmers's (2023) systematic review. Some studies, such as Crease and Blackledge (2010) and Li (2011), investigated translanguaging practices in Anglophone contexts but neither of these focused on interventions in HE settings. Rather, they looked into translanguaging classroom practices in heritage schools and students' interaction outside of classrooms (translanguaging spaces) in the UK, respectively. These findings show that studies investigating translanguaging in Anglophone HE settings are limited. Hence, it is reasonable to call for further research on translanguaging practice in this context.

Based on the discussion above, we suggest that translanguaging practices are potentially overlooked in Anglophone countries due to these contexts being excluded from traditionally defined EMI contexts, which have a heavy focus on non-English speaking countries. Thus, more studies and attention on translanguaging interventions in Anglophone countries are needed. To address this gap and shed light on how practitioners can apply pedagogical translanguaging in class, a practice review was conducted. A practice review is a unique type of research synthesis that aims to put "research into practice" through reviewing the findings of a pedagogical approach or practice from both academic and non-academic sources (Chong & Plonsky, 2023; Chong, 2025; also see Section 5 Overview of this book). This helps bridge the gap between research and practice and enhances research-informed practice or policy, making it suitable for our project. Our practice review aimed to classify translanguaging practices and to investigate them as a classroom intervention by reviewing and synthesising the current literature, both in academic and non-academic sources.

2. Methodological and ethical dilemmas/issues

The practice review is a unique type of research synthesis which, as the name suggests, has a focus on practice and pedagogy; it includes both academic and non-academic sources for analysis to put research into practice (Chong & Plonsky, 2023; Chong, 2025). The research team consists of a doctoral student at the University of St Andrews who is also a practitioner teaching at a UK university, her first supervisor who is an experienced researcher in multilingual settings, and the second supervisor who has a wealth of experience related to conducting research synthesis and reviews.

In 2025, Chong developed the Synthesis Methods and Reporting Tool (SMART)using the 'four dimensions of research syntheses quality: systematicity, transparency, accessibility, and reflexivity (STAR)' (p.7) (also see Chapter 4). This tool helps researchers to achieve best practice by providing clear guidelines on what authors have to consider before, during, and after conducting a research synthesis. It suggests that there are four important procedures in the preparation stage. Researchers should: (1) ensure their research is original; (2) form a research team including one expert on the topic and one expert on the methodology to ensure the reliability of the study; (3) name the project consistently, and (4) publish a protocol (Chong, 2025). Our research preparation process aligns with SMART (see Section 5 Overview of this book for detailed methodological stages for a practice review). Having done some background research on translanguaging (see detailed discussion in Section 1 of this chapter) and forming a research

team with experts on the topic and methodology, our research set off with a solid and positive base in March 2024. We then published a protocol in late August 2024 (see Kong et al., 2024).

Some may ask why it took a 'ready to go' group, a team with experts, almost half a year to prepare and publish a protocol. The preparation and research process, indeed, took longer than expected for several reasons. First of all, we are a busy team with different commitments alongside our work. However, the key obstacles were the methodological and ethical dilemmas that we faced during the process of preparing and conducting this practice review, which we will reflect upon in this section of our chapter.

In some cases, these dilemmas were anticipated, and in other cases, dilemmas revealed themselves as the research process progressed. Over the course of nine months (July 2024 – April 2025) we worked together on our practice review. Alongside this came our monthly supervision meetings, where we discussed issues related to the research. Partly to prepare for the current chapter, we decided to write reflections on the content of our meetings. We kept a reflective log on a Google Doc so that we could see each other's questions, concerns and thoughts as we went along. The reflective logs were documented shortly before July 2024 and after the protocol was published. They did not capture the whole process but the period towards the end of the protocol preparation process, and the ongoing research process.

2.1 Dilemma 1: Limited results found from both academic and non-academic sources

Firstly, in terms of methodology, in line with the definition of a practice review, we expanded our search to include practitioner magazines published in the last decade (from 2014 to 2024), including IATEFL's Voices magazine and English Teaching Professional. We aimed to find out how translanguaging practices are applied in classrooms by reviewing teachers' descriptions of those practices. We initially focused exclusively on the UK but we found very few articles that mentioned or described translanguaging practices in classrooms. Some of the articles mentioned 'using L1 in class' but none of them used the term 'translanguaging' or provided an explicit description of translanguaging as a classroom intervention. Additionally, most of the articles originated outside Anglophone countries and/or were not conducted in HE settings and for these reasons had to be excluded from our review. This initial result was unexpected and frustrating. We hoped and thought that we could gain insights into translanguaging practices by including and reviewing non-academic articles since the majority were written by practitioners and were practice focused. Nevertheless, only limited results were found,

which could be due to the gap between research and practice. Practitioners may not fully understand the concept of translanguaging and hence avoid using the term 'translanguaging' or the approach itself.

Limited relevant results were obtained not only from non-academic sources but also from the academic literature. At the pilot stage, we screened 209 academic articles but only three of them (Campbell & Tigan, 2022; Lin & Leung, 2023; Meletiadou, 2022) fitted our inclusion criteria, which was worrying. Ming Sum, the first author, documented this experience in one of her reflective logs (Extract 1).

Extract 1 After doing some research and screening, I found that it was quite likely that the selected articles would be about 10–20% of the searching results, which did not help because I was worried that it would not be significant to make a case or serve the purpose of the review.

(Ming Sum, Google Doc Reflection, July 2024)

We presented our initial findings and challenges at the British Association for Applied Linguistics (BAAL) conference in September 2024. After the conference, we wrote our reflections. Ming Sum reflected on the relevant but unclear descriptions of translanguaging interventions found in academic resources (Extract 2).

Extract 2 In order to have some initial results to present and talk about at the conference, I selected three relevant articles on translanguaging interventions. In fact, they are the only three from the 200 screened articles that are completely relevant. This further shows that we may have to widen our scope. Through the screening process and reading the three articles, I realised that even though the authors mentioned their interventions, they were usually vague and short. Students were usually 'encouraged' to use their repertoire or L1 rather than receiving clear instructions on when and how they could use them. This could be another hurdle of revealing what happens in a translanguaging classroom.

(Ming Sum, Google Doc Reflection, October 2024)

Based on these initial findings, a few questions were raised during this process: Are our context and inclusion criteria too specific and narrow, leading to these limited results? Are the limited discussions and descriptions of translanguaging practice due to a gap in the knowledge about how translanguaging can be applied in classrooms? Is this gap due to the lack of communication between researchers and practitioners? In fact, the later questions further demonstrate the need for doing this practice review. It is possible that the gap between research and practice and the lack of classification of translanguaging impede its use and discussion amongst practitioners. With these questions in mind, we reflected on our research scope and explored potential reasons and solutions.

2.2 Dilemma 2: Contested definitions of the Anglosphere

To reflect on our research scope, we had to look at our inclusion criteria and search terms. This led us to our second dilemma: an ethical consideration in defining the term 'Anglosphere' and determining which countries to include. On the one hand, we wanted to be inclusive and not focus only on 'inner circle' English-speaking countries such as the UK, the US, and Australia (Kachru, 1985, 2017). On the other hand, we needed to consider the scope, workload and feasibility of conducting the project.

Similar to conducting other research syntheses, we had to set 'research questions', 'inclusion and exclusion criteria', 'search strategies' and do 'screening' for a practice review (Chong, 2025, p. 10). In our review, we argue that translanguaging practices in Anglophone countries are worthy of investigation due to their high population of international students and the diversity of language exchange, yet it has been overlooked. Therefore, it is important to define the term 'Anglosphere' as this definition could affect the scope of our research. However, defining the 'Anglosphere' in terms of countries included and excluded, and indeed whether we should be looking at countries as discrete entities at all (Yuan, 2008), was a key challenge. In our broad view of English-medium higher education contexts, we wanted to focus on internationalised higher education in majority-English speaking countries (Mittelmeier et al., 2024). In the case of the UK, where we are both based, there is usually no hesitation in including it as part of the Anglosphere.

The term 'Anglosphere' was coined relatively recently, in 1995, by Neal Stephenson in his novel 'The diamond age' (Stephenson, 1995) but the term has gained prominence in political discourse to refer to a group of 'English-speaking nations that share a number of defining features: liberal market economies; the common law; parliamentary democracy; and a history of Protestantism' (Kenny & Pearce, 2018, p. 1). According to the Oxford English Dictionary (2025), the Anglosphere refers to 'English-speaking countries considered collectively; *esp.* the United Kingdom, the United States, Canada, Australia, and New Zealand (and sometimes other countries), regarded as a group with a shared historical and cultural heritage'. This definition highlights the countries that Kachru (1985, 2017) defined as 'inner circle' countries. The countries are often privileged in the field of TESOL and Applied Linguistics, thus, perpetuating 'native speakerism' (Holliday, 2006) and 'grand narratives' (Holliday, 2025) about the rightful ownership of English along with ideologies about standardised English and raciolinguistics (Alim et al., 2020). With 'the Anglosphere' most routinely being applied to the United Kingdom and its former settler dominions such as Canada, Australia and New Zealand (Kenny & Pearce, 2018), critics have questioned the boundaries of the term as applicable to today's postcolonial age of World Englishes and the

sociolinguistics of globalisation (Blommaert, 2010). Especially since the 1990s, the 'Anglosphere' has become an increasingly slippery and contested term.

As English is recognised not only as the property of so-labelled 'inner circle' countries but is often the first language, lingua franca, and/or the medium of instruction for many other countries globally, both former British colonies and otherwise, there is a need to broaden the definition of the Anglosphere beyond the countries named in the Oxford English Dictionary's definition. The complexities surrounding the term 'Anglosphere' made us question its value and relevance as a search term. Related concepts such as 'Anglophone' are equally contested (Research with International Students, n.d.). Although Anglophone generally describes English speakers or English-speaking nations, it can be applied in a plethora of ways that reflect ideologies and power constructs. In our study, when considering our search terms, we sought an inclusive definition of what constitutes the Anglosphere.

Given the difficulty of defining the Anglosphere and related terms, we discussed whether to narrow our focus to only one context. We conducted a pilot search of the UK, employing the search terms 'UK, England, Scotland, Wales, [and] Northern Ireland'. However, as discussed previously, our search on the UK higher education context alone did not produce enough articles to make our review a useful contribution. From 209 articles initially found, after screening, only three fit our criteria (see Extract 1). At this point, we realised the need to expand the search to our original criteria, which in turn brought us back to the dilemma over which countries to include.

When reflecting on this dilemma, we also expressed the emotions we encountered. Such emotions included frustration felt around ongoing discussions over contested terms, which resulted in delayed progress with our practice review, as seen in Extract 3:

Extract 3 Defining the scope of the review is always a challenge. Sarah and I have been going back-and-forth about the inclusion and exclusion criteria, e.g. whether we should include all Anglophone countries, such as Canada and the UK or simply the UK, and which level(s) of students should be included. It could be a bit daunting as it seemed that there was no progress, although Sarah reassured that it was good to have those in-depth discussions at the early stage. After doing some research and screening, I found that it was quite likely that the selected articles would be about 10–20% of the search results, which did not help because I was worried that it would not be significant to make a case or serve the purpose of the review.

(Ming Sum, Google Doc Reflection, October 2024)

Although our meetings about the two dilemmas felt tiresome and repetitive at times, they demonstrate an important sign and tip for conducting a research synthesis. That is, every decision made in the preparation stage is crucial and should be well-informed, cautious and well thought-through. This is because these decisions set as a base for the study, and they can affect the results of the research. Even a step as small as defining a key term could impact the data and the results obtained for the review. Thus, early-stage discussions within the research team are essential and invaluable.

After discussing our dilemmas, in the following section, we discuss our courses of action in relation to the methodological and ethical dilemmas we faced while conducting our practice review.

3. Courses of action

As discussed above, we faced two major dilemmas while preparing and conducting this practice review on translanguaging practice in Anglophone HE classrooms. They were: (1) limited results found from both academic and non-academic sources and (2) contested definitions of the key term 'Anglosphere'. These dilemmas led to several long discussions on potential solutions. In this section, we will talk about our process of overcoming the dilemmas and our courses of action.

3.1 Course of action on Dilemma 1: Limited results found from both academic and non-academic sources

Considering our first dilemma, having limited results from both academic and non-academic sources, we consulted the methodology expert in our team, Sin Wang Chong, to find out if we should be worried about the situation and if our resolutions were appropriate. During the meeting, we expressed our concerns and explored and discussed possible solutions (Extract 4).

Extract 4 Defining the scope of the review is always a challenge [...] However, after meeting with Sarah and Sin Wang today, I felt more positive. As an expert of research synthesis, Sin Wang suggested that even though the quantity of the studies we included in the end may not be extensive (though it would probably need to be more than ten studies), we could focus on quality than quantity. To enrich the results of our project, we could also collect other qualitative data by interviewing the author/practitioner of a relevant article and designing and sending out a survey to them to collect translanguaging materials. This gives me more insights and shows me different ways of collecting relevant and perhaps more in-depth data on a niche topic. (Ming Sum, Google Doc Reflection, November 2024)

After the meeting, it became clear that the quantity of data can be important as it can help enhance the credibility and make a stronger case for the study; however, the quality and depth of the data are equally valuable. Since we aim to investigate classroom intervention, having a small but high-quality in-depth data set could be more valuable than having a large but not entirely relevant set of data. Even though we could not directly resolve the dilemma that there was limited data in the literature, we could make the most of the existing data and expand on the data collection method. Therefore, we aimed to put more focus on the quality of the articles and decided that we could divide this project into two phases.

The first phase would be the current practice review, and the second phase would be a follow-up primary research project interviewing practitioners who claim that they have used translanguaging practices in class, especially the authors of the included articles in our practice review. Moreover, we could invite the authors to share their lesson plans and any materials that they have used in a translanguaging classroom. This can further bridge the gap between research and practice because this would provide a platform for dialogues between researchers and practitioners or even feedback from teachers to researchers. Through expanding the direction and drawing on a combination of data collection methods, we could potentially gain more insightful and fruitful data, despite the small data set.

Furthermore, to evaluate the quality of the data, we did a 'pilot screening' where we randomly selected and reviewed a few articles that fit our inclusion criteria. This allowed us to evaluate if the data was of high quality and suitable for in-depth analysis and to determine if the suggested solutions were indeed possible. It was a helpful and essential process when conducting a research synthesis (Extract 5 and Extract 6).

Extract 5 Reviewing articles with Sarah was also very helpful. This allows us to see how much and what data we can potentially get from the included articles and their quality. Thus, we can check if we are on the right track then decide if the scope is large enough. After the discussion, we think we can add more search terms (e.g. further education/postsecondary) and adjust our inclusion criteria (e.g. teachers' and students' perspective on the procedure of the practice) to enhance our search. The process and the outcome of the discussion shows that having a second researcher, esp. an expert in the field (Sarah), in the review process is crucial and critical. It not only helps enhance and ensure the quality of the review, but it also allows researchers to support each other both mentally and in action.

(Ming Sum, Google Doc Reflection, July 2024)

Extract 6 We continued today's meeting for an extra half hour after Sin Wang had left. We both felt that the project was starting to take a more structured shape. Ming Sum shared a selection of articles she felt matched the inclu-

> sion criteria for our practice review. This 'pilot screening' was very helpful as I could see that a range of studies applied, including studies on translanguaging amongst deaf and hearing students. Having recently returned from the Sociolinguistics Symposium (SS25) in Perth, Australia, I was reminded of Annelies Kusters's important work on 'crip linguistics' and 'linguistic caring'. Kusters's work discusses translanguaging beyond spoken 'languaging'. The reflection process is proving useful for sharing and refining our ideas. (Sarah, Google Doc Reflection, July 2024)

From Extracts 4, 5 and 6, we can see that having the opportunities to discuss issues raised, concerns and alternative solutions with a team of researchers and experts can be incredibly useful. These discussions can help to move the project forward and provoke thinking and re-thinking, hence, making the project more refined. Although we decided not to add search terms regarding education level in the end and chose to focus only on HE settings, we did expand our scope by adding other Anglophone countries to our search terms with the hope of potentially gaining more data. This leads us to our second dilemma solving process.

3.2 Course of action on Dilemma 2: Contested definitions of the key term 'Anglosphere'

Concerning our second dilemma, which was the difficulty of neatly defining the Anglosphere, we wanted to broaden the Oxford English dictionary definition to not only include the United Kingdom, the United States, Canada, Australia, and New Zealand, but other countries too, as reflective of World Englishes in today's era of the sociolinguistics of globalisation (Blommaert, 2010). Ming Sum looked at the way UK university websites defined 'majority English-speaking countries' for the purpose of exempting citizens from these countries from having to take an English proficiency university entrance exam. Often, such lists were based on the UK Government (2024)'s definition of 'majority English-speaking countries', as explained in Extract 7.

Extract 7 I looked at the definition(s) of the Anglosphere and what can be considered as an Anglophone country. I found a list of 19 countries proposed by the UK Government (2024) which UK universities (e.g. Swansea and Liverpool) call it 'Majority English-speaking countries'. Students who are from these countries are not required to prove their knowledge of English. I believe that this definition suits us well since it is set from the UK government for universities. I am glad that Sarah and I both agreed on this and the direction of expansion. I hope that this can help us move forward and enhance our study. (Ming Sum, Google Doc reflection, October 2024)

In Extract 7, there is a sense of moving toward a solution to the dilemma, and relief which accompanies this. Although the UK Government's (2024) list seemed like a good definition to use and would produce a greater number of relevant articles, when discussing it further between us and with colleagues, we saw the need to edit the search terms further due to a British bias toward its overseas territories and British postcolonial contexts. We discussed the list and agreed that the list would also need to be critiqued in our review, as some countries we expected to see on the list (e.g., South Africa and Nigeria) were not there. After agreeing to expand the scope using the UK Government's (2024) definition of 'majority English Speaking countries', we did individual searches for each country with our main search string on different databases. Perhaps not surprisingly, due to the legacy of colonialism, the so-labelled 'inner circle countries' such as the UK, the US, Canada and Australia were the most common contexts for 'translanguaging in practice' research. Other countries on the list had limited or no results, such as Barbados and Guyana. The same pattern emerged when we searched for each British overseas territory. Such results speak to uneven power relations in terms of internationalisation in higher education, research coverage, publishing, and visibility (Piller, 2022). We reflected on how the process of working on the practice review continued to be educational in this sense, as seen in Extract 8.

Extract 8 We were surprised that some countries were missing from the list. On reflection, the list is 'UK-centric' as the countries were all former colonies or territories of the UK. This reiterated the contested nature of 'Anglosphere' and 'Anglophone' for us, and it makes choosing inclusion criteria 'slippery'. The notion of 'contested terms' is one that occurs over and over again across many research areas. Another contested term is 'native speaker', for example. (Sarah, Google Doc reflection, November 2024)

Thinking about how these insights affect our study and courses of action, we recognised that it was important to openly discuss how we selected our search string and provide a detailed and transparent description for every step of the process.

Through discussion and reflection, the key dilemmas were resolved. We decided to focus on quality rather than the quantity of the data and expand our data collection methods, for instance, by interviewing practitioners, thus branching into a second phase of the study. Moreover, we expanded the scope for our research based on a list of Anglophone countries suggested by the UK Government (2024), while acknowledging that the terms 'Anglosphere' and 'Anglophone' were contested and that the list we chose to use was not fully inclusive.

Although the process of preparation was longer than we expected and sometimes felt repetitive, we valued those conversations and believe that this process will serve as a good foundation and provide a clearer direction for our study.

Using the revised scope, in the second pilot screening, our inclusion rate increased from 10–20% to nearly 30%. This shows that the solution was effective, and the early-stage discussions were worth having.

4. Alternative courses of action

During the exploration and discussion period, apart from the above actions, other alternatives were suggested. Our two dilemmas, a small sample and a contested definition of 'Anglosphere', are closely related. Thus, our discussion on alternative courses of action was mainly focused on expanding our scope, such as adding more search terms and widening the contexts.

As mentioned in Section 3, we had considered adding search terms regarding the education contexts, i.e. adding 'post-secondary' and 'further education' to the search string 'universit*' and 'higher education'. However, the levels of study and qualifications of the two suggested terms could be different from those at universities and higher education settings, depending on countries. As we argue, more research and attention should be directed toward HE settings, since they have been overlooked (as discussed in Section 1); post-secondary and further education studies may not fit our purposes. Equally, to expand our search but stay focused on an Anglophone context, we could have focused only on the UK context (or another single context), but across all levels of education. Doing the latter would have provided a wider range and larger number of articles on translanguaging practice for our review. Nevertheless, our research challenges the traditional EMI definition and supports the framework of 'EMEMUS' (Dafouz & Smit, 2016) which required us to focus specifically on higher education. We, therefore, chose to expand our scope by adding new search terms regarding countries only.

When considering adding search terms regarding education contexts and expanding the research scope to all education levels in the UK, we also had a brief discussion on using AI to help with the review. This was because taking those alternative courses of action would have led to an extensive amount of data. This idea was inspired by another peer researcher who used AI to screen 10% of his data (around 1000 papers) to save time on the screening process. A reflection was documented after this discussion (see Extract 9).

Extract 9 In addition, we talked about potential AI use for data screening since I was worried that the results would be unmanageable (e.g. 2500 or above). Sin Wang has no objection to AI use but again, we have to provide a reasonable justification for its use, especially if we would like to submit our work to the top journals. This discussion reminds me of how important rigorous methodology is in good-quality research at this level, especially when con-

> ducting a review. It really takes time and effort to search and synthesise the data. I am relieved that our results are less than 2000 which is still a lot but manageable. (Ming Sum, Google Doc reflection, November 2024)

This discussion reminded us that the basis of conducting research is to ensure the decisions we make are justifiable, evidence-based, and rigorous. It also prompted us to think about a potential home for this practice review.

When reflecting on alternative courses of action for the second dilemma, we could have focused on a specific set of named countries which shared similar traits and avoided the terms 'Anglosphere and Anglophone' altogether. Further issues exist in relation to the term 'Anglophone' as it is often used as a binary with 'non-Anglophone', which implies that 'Anglophone' is the norm and 'non-Anglophone' is a deficit, due to the prefix of 'non' (Research with International Students, n.d.). Considering these arguments, avoiding the term in our research could have been an alternative action.

Ideally, if we had enough capacity, we could have adopted a wider definition of Anglophone and included studies from all levels of study in every country with English as its official language. This way, these dilemmas would have been resolved, and a more holistic review could have been produced. Nevertheless, it should be acknowledged that there might well have been other dilemmas that would have arisen.

5. Conclusion

In this chapter, we discussed the rationale and reflected on the process of conducting a research synthesis on translanguaging practices in Anglophone classrooms. This project is a practice review aiming to find out (1) what is classified as translanguaging practices? and (2) what does translanguaging look like in classrooms? We argue that despite the wide discussion on translanguaging, the meanings, classification, and the application of translanguaging are still vague and inconsistent (Treffers-Daller, 2025). This could create difficulties for researchers to conduct research on translanguaging practices and practitioners to apply them in classrooms. Therefore, there is a need to review and synthesise translanguaging research, both academic and non-academic texts (Chong & Plonsky, 2023) (i.e. practitioner magazines) in Anglophone HE settings to bridge the gap between research and practice.

Our research team consisted of three researchers, a doctoral student who is also a practitioner teaching at a UK university, an expert on the topic, and an expert on the methodology (Chong, 2025). To document our collaborative

process, mainly for the purpose of this chapter, two of the team members (the authors) wrote short reflections and kept them as reflective logs after each supervision meeting (between July 2024 and April 2025). Some of them were presented in this chapter to support our descriptions of this reflective process.

During our reflection, we identified and discussed our two key methodological and ethical dilemmas, a small data set and difficulties when defining the key term 'Anglosphere', and the corresponding and alternative courses of action. After some further research and discussions, to gain more in-depth and high-quality data, we decided to divide our practice review into two phases: (1) the current practice review synthesising existing literature; (2) a primary research interviewing practitioners with experience in using translanguaging in classrooms. Furthermore, we expanded the scope using a list of 19 countries, which some UK universities named as 'majority English speaking countries', proposed by the UK Government in 2024. As the definition is from an official source, it is a good base, which we were able to adapt to include more studies, enhancing our dataset due to this new scope. This course of action helped address the two dilemmas.

Through this collaborative and reflective process, we realised that the process of conducting a research synthesis can be complex, especially at the preparation stage. It is because it involves a lot of thinking and procedures such as researching, piloting, registering the review, and refining search terms and the scope or even topics. This process can be daunting and frustrating as it seems repetitive. However, each little step and decision in this process is critical and invaluable. Each decision can impact the feasibility, the following stages, and the results of the project. Therefore, it is important to have a research team consisting of members who have different expertise and experience and who come from different professional roles (i.e. researcher and practitioner), especially when conducting practice-focused research. Apart from ensuring the credibility and reliability of the research, team members can provide not only moral support, but also concrete and useful suggestions when concerns are raised. Each of our team meetings enhanced the research process, such as learning about different ways of collecting data and discussing the essential elements of a piece of high-quality research, e.g. a rigorous methodology. Our discussion also offered new perspectives and prompted us to rethink and reflect on our decisions to make appropriate refinements. Keeping a reflective log also helped us to clarify thoughts and keep track of our progress, which can be encouraging when looking back and seeing how far we had come.

Overall, conducting a research synthesis requires hard work, persistence, detailed thinking, and teamwork. Nevertheless, the sense of achievement when seeing the output and helping to bridge the gap between research and practice in teaching and learning is precious and meaningful.

6. Questions for discussion

I. What makes you interested in doing a practice review?
II. The authors discuss their dilemmas when conducting a practice review and the benefits of working together as a research team. Reflecting on your experience in conducting a research synthesis, what are the benefits and challenges of working as part of a carefully selected research team, especially a team with practitioner members?
III. The authors discuss contested terms (Anglosphere and Anglophone) and how such terms affect ethics around developing an inclusive search string. What are some contested terms or Western-centric terms related to your research that may need to be critiqued at the stage of deciding on search terms? How might this affect the review process?
IV. The authors of this chapter made use of a research log to document and reflect on the review process. What are other ways to facilitate reflective dialogue when conducting a research synthesis?

Further reading

I. Bonacina-Pugh, F., da Costa Cabral, I., & Huang, J. (2021). Translanguaging in education. *Language Teaching*, *54*(4), 439–471.

This state-of-the-art article reviews the existing definitions and concepts of translanguaging, explores its potential impact and significance in different educational contexts, and offers direction for future research. It provides an overview of the theoretical and practical discussion of translanguaging in the field clearly and systematically.

II. Chong, S.W. (2025). Synthesis Methods and Reporting Tool (SMART) for research syntheses in Applied Linguistics. *Research Synthesis in Applied Linguistics*, *1*(1), 17–38.

This article gives an overview of different types of research synthesis in Applied Linguistics and a step-by-step guide on how to conduct a research synthesis using SMART. It is clear and easy to follow. It also provides at least one example for each type of research synthesis, which allows the reader to contextualise what was introduced in the paper and how it could be done in practice.

III. Chong, S.W., Oxley, E., Negrea, V., Bond, M., Liu, Q., & Kong, M.S. (2024). *Teacher recruitment and retention in schools in socioeconomically disadvantaged areas in England: review of practice*. Education Endowment Foundation. https://educationendowmentfoundation.org.uk/education-evidence/evidence-reviews/recruitment-and-retention-in-disadvantaged-schools

This is a practice review example looking into teacher recruitment and retention strategies in socioeconomically disadvantaged areas in England through reviewing academic sources and teacher job postings. The researchers also sent out a survey to teachers. The findings of the review and survey were compared and discussed in order to get a comprehensive view on the

topic to inform future policy-making. This is a vivid example of research into practice and research-informed practice.

IV. Research with International Students (n.d.). *Anti-Glossary" of contested terms*. https://researchintlstudents.com/anti-glossary/

This academic blog post provides an important discussion on the questioning of problematic terms for meaningful communication and critique. The academic blog post looks at how researchers and practitioners can engage reflexively with terminology and think carefully about and address the assumptions that underpin contested terminologies.

AI use declaration statement

The authors confirm that no AI tools were used in the preparation, writing, formatting, and editing of this work.

References

Alim, H.S., Rickford, J.R., & Ball, A.F. (Eds.). (2020). *Raciolinguistics: How language shapes our ideas about race*. Oxford University Press.

Blommaert, J. (2010). *The sociolinguistic of globalization*. Cambridge University Press.

Bonacina-Pugh, F., da Costa Cabral, I., & Huang, J. (2021). Translanguaging in education. *Language Teaching*, *54*(4), 439–471.

Campbell, M., & Tigan, A. (2022). Translanguaging and product-oriented drama: An integrated pedagogical approach for language learning and literacy development. *International Journal of Bilingual Education and Bilingualism*, *25*(10), 3745–3757.

Caruso, E. (2018). Translanguaging in higher education: Using several languages for the analysis of academic content in the teaching and learning process. *Language Learning in Higher Education*, *8*(1), 65–90.

Chang, S.Y. (2019). Beyond the English box: Constructing and communicating knowledge through translingual practices in the higher education classroom. *English Teaching & Learning*, *43*(1), 23–40.

Chong, S.W. (2025). Synthesis Methods and Reporting Tool (SMART) for research syntheses in Applied Linguistics. *Research Synthesis in Applied Linguistics*, *1*(1), 17–38.

Chong, S.W., & Plonsky, L. (2023). A typology of secondary research in Applied Linguistics. *Applied Linguistics Review*, *15*(4), 1569–1594.

Creese, A., & Blackledge, A. (2010). Translanguaging in the bilingual classroom: A pedagogy for learning and teaching? *The Modern Language Journal*, *94*(1), 103–115.

Dafouz, E., & Smit, U. (2016). Towards a dynamic conceptual framework for English-medium education in multilingual university settings. *Applied Linguistics*, *37*(3), 397–415.

Dafouz, E., & Smit, U. (2020). *ROAD-MAPPING English medium education in the internationalized university*. Palgrave Macmillan.

García, O. (2012). Theorizing translanguaging for educators. In C. Celic & K. Seltzer (Eds.), *Translanguaging in the bilingual classroom: A pedagogy for learning and teaching?* (pp.1–6). CUNY-NYSIEB, The Graduate Center.

García, O., & Lin, A. M. (2017). Translanguaging in bilingual education. In O. Garcia, A. Lin, & S. May (Eds.), *Bilingual and multilingual education* (pp.1–14). Springer Nature.

García, O., & Otheguy, R. (2020). Plurilingualism and translanguaging: Commonalities and divergences. *International Journal of Bilingual Education and Bilingualism*, *23*(1), 17–35.

Holliday, A. (2006). Native-speakerism. *ELT Journal*, *60*, 385–387.

Hopkyns, S. (2022). Translanguaging in Gulf English-medium instruction higher education. In F.A. Rashdi & S.R. Sandhya (Eds), *Language and identity in the Arab world* (pp.76–92). Routledge.

Huang, X., & Chalmers, H. (2023). Implementation and effects of pedagogical translanguaging in EFL classrooms: A systematic review. *Languages*, *8*(3), 194.

Kachru, B. (1985). *Standards, codification and sociolinguistic realism: English language in the outer circle*. In R. Quirk & H. Widdowson (Eds.), *English in the world: Teaching and learning the language and literatures* (p. 11–36). Cambridge University Press.

Kachru, B. B. (2017). *World Englishes and culture wars*. Cambridge University Press.

Kenny, M., & Pearce, N. (2018, May 14). *In the shadows of empire: How the Anglosphere dream lives on*. https://blogs.bath.ac.uk/iprblog/2018/05/14/in-the-shadows-of-empire-how-the-anglosphere-dream-lives-on/

Kong, M. S., Hopkyns, S., & Chong, S. W. (2024). *A protocol for a practice review of translanguaging in UK classrooms*. International Database of Education Systematic Reviews. https://idesr.org/article/IDESR000158

Kusters, A., Spotti, M., Swanwick, R., & Tapio, E. (2017). Beyond languages, beyond modalities: Transforming the study of semiotic repertoires. *International Journal of Multilingualism*, *14*(3), 219–232.

Lin, S., & Leung, A. (2023). ESL classroom interactions in a translanguaging space. *Applied Linguistics Review*, *15*(6), 2397–2425.

Li, W. (2011). Moment analysis and translanguaging space: Discursive construction of identities by multilingual Chinese youth in Britain. *Journal of Pragmatics*, *43*(5), 1222–1235.

Llanes, À., & Cots, J. M. (2022). Measuring the impact of translanguaging in TESOL: A plurilingual approach to ESP. *International Journal of Multilingualism*, *19*(4), 523–538.

Meletiadou, E. (2022). The utilisation of peer-assisted learning/mentoring and translanguaging in higher education. *IAFOR Journal of Education*, *10*(1), 135–154.

Mittelmeier, J., Lomer, S., Al-Furqani, S., & Huang, D. (2024). Developing meaningful internationalisation that impacts students' outcomes in higher education: A scoping review of the literature 2011–2022. *Journal of Studies in International Education*, *28*(4), 503–525.

Oxford English Dictionary (2025). Anglosphere. https://www.oed.com/dictionary/anglosphere_n?tl=true#:~:text=To%20be%20part%20of%20the,core%20of%20English%2Dspeaking%20cultures

Pecorari, D., & Malmström, H. (2018). At the crossroads of TESOL and English Medium Instruction. *TESOL Quarterly*, *52*, 497–515.

Piller, I. (2022, May 11). How to challenge Anglocentricity in academic publishing. *Language on the Move*. https://www.languageonthemove.com/how-to-challenge-anglocentricity-in-academic-publishing/

Prilutskaya, M. (2021). Examining pedagogical translanguaging: A systematic review of the literature. *Languages*, *6*(4), 180.

Research with International Students (n.d.). *Anti-Glossary" of contested terms*. https://researchintlstudents.com/anti-glossary/

Stephenson, N. (1995). *The diamond age*. Bantam Spectra.

Treffers-Daller, J. (2025). Translanguaging: What is besides smoke and mirrors? *Linguistic Approaches to Bilingualism*, *15*(1), 1–26.

UK Government. (2024). *Student Visa*. Retrieved on 20 July 2025 from https://www.gov.uk/student-visa/knowledge-of-english

Vogel, S., & García, O. (2017). Translanguaging. *Oxford research encyclopedia of education*. Retrieved on 27 April 2025 from https://oxfordre.com/education/view/10.1093/acrefore/9780190264093.001.0001/acrefore-9780190264093-e-181.

Wang, W., & Curdt-Christiansen, X. L. (2019). Translanguaging in a Chinese–English bilingual education programme: A university-classroom ethnography. *International Journal of Bilingual Education and Bilingualism*, *22*(3), 322–337.

Yuan, M. (2008). Discrete versus continuous phenomena. In *Encyclopedia of geographic information science* (pp. 112–114). Sage.

SECTION 6

Scoping review

Section overview

What is a scoping review?

Mara Fuertes Gutiérrez & Inma Álvarez
The Open University, United Kingdom

1. Definition

Scoping reviews (ScR), also called "mapping reviews" or "scoping studies", represent a systematic, rigorous and transparent approach to document existing research on a given domain (Pham et al., 2014), offering extraordinary value to "preliminarily survey qualitative and quantitative evidence on a niche or emergent topic" (Chong, 2025, p.7). Within the broader landscape of research research synthesis, ScR are categorised as a type of "big picture" review which is exploratory (Campbell et al., 2023), alongside bibliometric reviews (Chong, 2025). This type of review plays a crucial role in summarising knowledge from published and unpublished studies. Rather than synthetising findings, ScR focus on mapping available research evidence with the aims of informing practices and suggesting future lines of research (Arksey & O' Malley, 2005). They can also help assess whether conducting a systematic review is needed or justified (Tricco et al., 2018). By providing analytical in-depth overviews of a diversity of publications with the purpose of identifying knowledge gaps, types of available evidence, key characteristics of concepts or research, scoping reviews are essential in supporting research-lead teaching practices and constitute a useful tool for researchers working in emerging and evolving topics in Applied Linguistics.

Methodological stages

Scoping reviews are undertaken by a team of researchers to increase reliability. A consultation with stakeholders is recommended throughout the process of the review as well as using an intercoderreliability test to guarantee consistent coding.

There is considerable advice available for researchers regarding methodological frameworks for conducting scoping reviews. An initial framework was proposed by Arksey and O'Malley (2005), which was later refined and expanded by

https://doi.org/10.1075/rmal.17.11gut

Aromataris et al. (2024). The Preferred Reporting Items for Systematic Reviews (PRISMA) has an extension for scoping reviews, the PRISMA-ScR, published in 2018 (Tricco et al., 2018; see also Peters et al., 2020). Based on this previous guidance, the following steps are recommended when conducting scoping reviews in Applied Linguistics:

1. **Definition of objectives and questions.** Given the exploratory nature of ScR, it is essential to formulate review questions and objectives with a broad scope. To achieve this, it is recommended to use the PCC (Population, Concept, Context) framework, as it clearly defines key elements. For Applied Linguistics, population might include speakers, learners, pre- and in-service teachers, teacher trainers, parents, curriculum developers or policymakers; concept could be a range of topics such as a teaching practice, an aspect of learning, or curriculum design; and context might feature geographical settings, educational level, or languages to be covered in the ScR.
2. **Development of eligibility criteria**. Develop inclusion and exclusion criteria which aligns with objectives and questions, providing a rationale for each. In the case of Applied Linguistics, inclusion criteria usually contemplate time period, or language of publication. Importantly, given the large amount of grey literature in the field, establishing the study types that will be included in the ScR (empirical, theoretical, literature reviews, grey literature) is particularly relevant. Moreover, given the wide variety of research methods used in Applied Linguistics (qualitative, quantitative, mixed-methods), reviewers might also want to consider research design as part of their inclusion/exclusion criteria, depending on the aims of the review. Frequent exclusion criteria tend to include studies focusing on first language acquisition or editorials and opinion pieces.
3. **Plan of search strategy**. Define a comprehensive search approach, extraction of data and approach to presentation of evidence. This stage considers which keywords are relevant to the topic.
4. **Search of evidence**. Interrogate relevant subject-specific databases to find the evidence. In Applied Linguistics, the use of a single database such as Web of Science or Scopus might mean important contributions are left out, especially those written in languages other than English; thus it is advisable to use complementary databases such as ERIC or the Linguistics and Language Behavior Abstracts (LLBA).
5. **Identification and selection of sources of evidence**. Two screening phases are required: screening of titles and abstracts of articles by two researchers using the criteria; and screening of full-texts deem relevant.

6. **Data extraction**. Select eligible sources of evidence and present details of the characteristics of the data found.
7. **Analysis of the evidence**. Interpret the findings.
8. **Presentation of the results**. Present main results answering the questions and objectives set up for the study.
9. **Summary of the evidence**. Present a descriptive synthesis, draw conclusions, explain potential implications and inform future research.

In addition to this process, researchers should engage with the Synthesis Methods and Reporting Tool (SMART) for RS in Applied Linguistics (Chong, 2025) and use it as guide to design the methodological approach and reporting strategy when conducting ScR.

Software/resources

Researchers conducting ScR can access numerous resources to support their work at various stages. We have identified five useful tools, but for a comprehensive list, see, for instance, Cierco Jiménez et al. (2022) or the list provided by RSAL (n.d.). While the role of artificial intelligence in ScR is still emerging, some tools, such as EPPI Reviewer, are already experimenting with incorporating tools such as ChatGPT.

1. Databases for search strategy: The Open University Library (United Kingdom)
   It is recommended that a minimum of 3–4 databases should be used to conduct a ScR. The decision of which databases should be included would depend on the topic and the inclusion and exclusion criteria. This extensive list of databases (https://shorturl.at/5sMrb) provided by The Open University supports researchers in deciding which may be the most convenient. Please note that this portal to databases does not guarantee access to all databases or the research, but provides a description of their content and a link if they are open access.
2. Referencing tools and data management: Zotero
   This open access reference management software assists with archiving the literature, supporting in-text citations and building reference lists. It is available in a desktop and online version and works well for team projects since it allows to host shared libraries and facilitates collaboration.
3. Data extraction: EPPI-Reviewer
   This software supports all stages of the methodological work involved in conducing ScR, from data extraction to coding and screening. It also allows to work collaboratively.

4. Reporting: OASIS
   This tool helps authors to disseminate their reviews, as it offers one-page summaries of published research on language-related topics. Practitioners are encouraged to engage with OASIS (https://www.oasis-database.org/.), which ultimately contributes to the use of ScR to inform language teaching. Another platform that hosts accessible summaries of ScR is TESOLgraphics, which produces infographic summaries of research syntheses (www.tesolgraphics.com).
5. Quality assurance: AMSTAR 2
   The Assessing the Methodological Quality of Systematic Reviews (AMSTAR 2 https://amstar.ca/Amstar_Checklist.php) checklist can be used to measure the quality of ScR.

Examples

The following five publications serve as good examples of scoping reviews. Their selection is based on several criteria: they are published as open access, cover a range of topics and complement each other by illustrating different aspects of best practices. We have also taken into account publication date, and therefore these studies follow the latest methodological advancements in conducting ScR:

Chong, S. W., & Reinders, H. (2025). Autonomy of English language learners: A scoping review of research and practice. *Language Teaching Research*, *29*(2), 607–632.

This review exemplifies how ScR can be used to explore research available on a core concept and to illustrate the dual focus of ScR on research and practice. Using Chong and Plonsky's (2021) methodological framework, it offers a very useful, comprehensive overview of included studies (pp. 613–615).

Hiver, P., Al-Hoorie, A. H., & Evans, R. (2022). Complex dynamic systems theory in language learning: A scoping review of 25 years of research. *Studies in Second Language Acquisition*, *44*(4), 913–941.

This review serves as an example of an investigation on how a new theoretical paradigm has been researched reporting on methodological approaches, key contributions and areas of improvement. An in-depth explanation of inclusion criteria and a detailed description of the coding process are provided, along with a clear account of how intercoder agreement was achieved. Reporting clearly follows research questions established.

Ramezanzadeh, A.-M., & Woore, R. (2023). *A scoping review into the teaching and learning of Arabic as an additional language*. Qatar Foundation International. https://shorturl.at/tv2ck

This review, commissioned by a non-profit organisation and published on their website, adopts a report-style format, highlighting the role of ScR in Applied Linguistics beyond academia. It demonstrates the importance of developing and registering a review protocol (https://idesr.org/article/IDESR000054). The use of visual representations of findings is also a key strength and includes the names of an Expert Advisory Panel.

Siegel, A., Vance, M., & Nilsson, D. (2024). Military English language education: A scoping review of 30 years of research. *Innovation in Language Learning and Teaching*, *19*(5), 433–451.

This review includes a strong, well-justified rationale for using a scoping review approach to explore an under-researched topic. Although it does not calculate intercoder reliability, it offers a thoughtful alternative by employing a collaborative screening process.

Weisleder, A., Friend, M., Tsui, A. S. M., & Marchman, V. A. (2024). Using parent report to measure vocabulary in young bilingual children: A scoping review. *Language Learning*, *74*(2), 468–505.

This review provides in-depth details of a search strategy and inclusion and exclusion criteria. It has a specific section that clearly states the limitations of the review.

AI use declaration statement

The authors confirm that no AI tools were used in the preparation, writing, formatting, and editing of this work.

References

Arksey, H., & O'Malley, L. (2005). Scoping studies: Towards a methodological framework. *International Journal of Social Research Methodology*, *8*(1), 19–32.

Aromataris, E., Lockwood, C., Porritt, K., Pilla, B., & Jordan, Z. (Eds.). (2024). *JBI manual for evidence synthesis*. JBI. https://synthesismanual.jbi.global.

Campbell, F., Tricco, A. C., Munn, Z., Pollock, D., Saran, A., Sutton, A., White, H., & Khalil, H. (2023). Mapping reviews, scoping reviews, and evidence and gap maps (EGMs): The same but different- the "Big Picture" review family. *Systematic Reviews*, *12*(1), 45–48.

Chong, S. W. (2025). Synthesis Methods and Reporting Tool (SMART) for research syntheses in Applied Linguistics. *Research Synthesis in Applied Linguistics*, *1*(1), 17–38.

Chong, S. W., & Plonsky, L. (2021). A primer on qualitative research synthesis in TESOL. *TESOL Quarterly*, *55*(3), 1024–1034.

Chong, S. W. & Plonsky, L. (2024). A typology of secondary research in Applied Linguistics. *Applied Linguistics Review*, *15*(4), 1569–1594.

Cierco Jimenez, R., Lee, T., Rosillo, N., Cordova, R., Cree, I. A., Gonzalez, A., & Indave Ruiz, B. I. (2022). Machine learning computational tools to assist the performance of systematic reviews: A mapping review. *BMC Medical Research Methodology*, *22*, 322.

Munn, Z., Peters, M. D. J., Stern, C., Tufanaru, C., McArthur, A. & Aromataris, E. (2018). Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. *BMC Medical Research Methodology*, *18*, 143.

Peters, M. D. J., Marnie, C., Tricco, A. C., Pollock, D., Munn, Z., Alexander, L., McInerney, P., Godfrey, C. M., & Khalil, H. (2020). Updated methodological guidance for the conduct of scoping reviews. *JBI Evidence Synthesis*, *18*(10), 2119–2126.

Resources for Conducting Research Syntheses. Weebly. (n.d.). https://researchsynthesis.weebly.com/resources-for-conducting-research-syntheses.html

Pham, M. T., Rajić, A., Greig, J. D., Sargeant, J. M., Papadopoulos, A., & McEwen, S. A. (2014). A scoping review of scoping reviews: Advancing the approach and enhancing the consistency. *Research Synthesis Methods*, *5*(4), 371–85.

Tricco, A. C., Lillie, E., Zarin, W., O'Brien, K. K., Colquhoun, H., Levac, D. et al. (2018). PRISMA extension for scoping reviews (PRISMA-ScR): Checklist and explanation. *Annals of Internal Medicine*, *169*, 467–473.

CHAPTER 11

Reflecting on scoping review methodology with a language testing example

Sha Liu
British Council

This chapter provides a reflective analysis of conducting a scoping review on L2 learners' cognitive processes in language testing as a methodological case study (for the published review protocol, see Chan & Chong, 2023). Guided by Arksey and O'Malley's (2005) framework, the chapter explores both anticipated and emergent methodological and ethical challenges, the courses of action taken, and alternative pathways not pursued. Central themes include the tension between comprehensiveness and feasibility, and the role of AI in research synthesis. Drawing on the team's direct experience, the chapter distinguishes between mechanical tasks suitable for AI automation and interpretive tasks requiring human expertise, concluding with key lessons for navigating similar challenges in research synthesis.

Keywords: scoping review methodology, language testing, L2 cognitive processes, AI-assisted research synthesis, methodological reflection

1. Background

Research into second language (L2) test-takers' cognitive processes has evolved rapidly since Weir's (2005) socio-cognitive framework foregrounded cognitive validity as an integral component of test validation. Following this paradigm shift, the past two decades have witnessed proliferating studies across all four language skills. These studies employ increasingly diverse methodologies, ranging from traditional think-aloud protocols (Stratman & Hamp-Lyons, 1994) to more recent approaches including stimulated recall (Brunfaut & McCray, 2015), keystroke logging (Chan, 2017), eye-tracking (Yu, He, & Isaacs, 2017), and even neuroimaging techniques like functional near-infrared spectroscopy (fNIRS, Zhai & Aryadoust, 2022).

While these methodological innovations have yielded valuable insights into test-taking behavior, they have outpaced conceptual consolidation. Three interre-

https://doi.org/10.1075/rmal.17.11aliu
Available under the CC BY-NC-ND 4.0 license.© 2026 John Benjamins Publishing Company

lated problems impede cumulative interpretation. First, researchers employ key terms such as process, strategy, and subskill inconsistently and often interchangeably, lacking clear operational definitions. A systematic review of L2 listening assessment by Aryadoust and Luo (2023) revealed "unresolved issues concerning the definition and operationalization" of constructs, particularly the confusion between subskills and processes (p. 375). Second, no common analytical framework exists for handling the increasingly fine-grained data produced by new technologies. Third, attempts to align observed behavior with theoretical accounts of language-use cognition remain tentative.

To navigate this evolving research domain, we adopted a scoping review methodology, one of the 13 evidence synthesis approaches currently recognized in Applied Linguistics (Chong & Plonsky, 2024). For a comprehensive introduction to scoping reviews within the broader landscape of research synthesis, see Fuertes Gutiérrez and Álvarez (Section Overview, this volume). Scoping reviews are "exploratory" (Chong, 2025, p. 5) and differ fundamentally from systematic reviews that answer narrowly focused questions. Instead, they enable researchers to map the extent, range, and nature of evidence, identify gaps, and discern trends in emerging or conceptually expansive fields (Arksey & O'Malley, 2005; Levac et al., 2010; Peters et al., 2020). This mapping approach is particularly well-suited for research areas that are broad in scope, interdisciplinary, and have undergone significant paradigm shifts, as demonstrated by Terrier et al. (2025) in their scoping review of 25 years of L2 pronunciation research. Our review topic exhibits precisely these characteristics: L2 cognitive process research spans multiple language skills and methodological traditions, draws on frameworks from both language testing and cognitive psychology, and has undergone a paradigm shift since Weir's (2005) socio-cognitive framework. These features, combined with the conceptual ambiguity surrounding key constructs (discussed in Section 2.1), made a scoping review the most appropriate synthesis approach for mapping this evolving domain (Munn et al., 2018).

Our review aimed to address four dimensions of the field: the language skills investigated, assessment contexts and participant profiles, research methodologies employed, and findings regarding test-takers' cognitive processes. These research questions balanced breadth with operational feasibility, capturing conceptual variety while maintaining manageable boundaries, which is a key consideration in scoping review design (Arksey & O'Malley, 2005). Addressing these questions promised to synthesize methodological developments, identify limitations, and inform future research agendas. This mapping is particularly timely given the emergence of AI-based multimodal assessment tasks that require test-takers to orchestrate multiple language skills simultaneously across textual, audio, and visual inputs. Whether these complex tasks activate their intended cognitive

processes remains unknown, underscoring the urgent need to investigate their construct validity.

Figure 1 illustrates the six-stage framework (Arksey & O'Malley, 2005) that guided our review process, providing structure for the methodological reflections presented throughout this chapter.

1. Identifying the research question: Establishing review boundaries via Population-Concept-Context parameters (Peters et al., 2020).
2. Identifying relevant studies: Systematic searching across databases, reference lists, and grey literature.
3. Selecting studies: Two-phase screening (title/abstract; full-text) applying predetermined inclusion/exclusion criteria.
4. Charting the data: Standardized extraction using piloted forms for descriptive data collection.
5. Collating, summarizing and reporting results: Numerical and thematic synthesis to map literature patterns and gaps.
6. Consulting stakeholders (optional): External validation to enhance interpretation (not undertaken in this review).

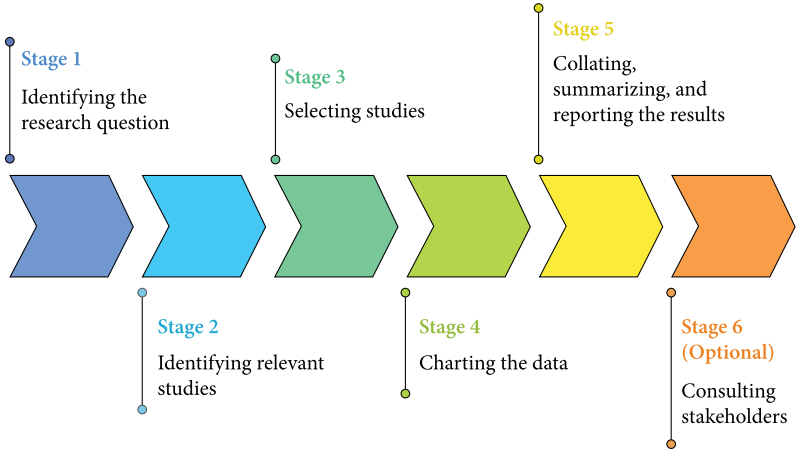


Figure 1. Six-stage scoping review framework (adapted from Arksey & O'Malley, 2005)

2. Methodological and ethical dilemmas: Anticipated and emergent challenges

Conducting our scoping review revealed two categories of challenges: those anticipated during protocol development based on methodological literature (Section 2.1), and those emerging unexpectedly during implementation (Section 2.2).

2.1 Anticipated challenges: Navigating foreseen complexities

Drawing upon the published review protocol (Chan & Chong, 2023) and the research literature, we identified several key challenges. Primary among these was the lack of consensus surrounding our core constructs: cognitive processes and language testing. Researchers employ diverse terms such as process, strategy, subskill, and function to describe test-taker cognition, while test, exam, and assessment are used interchangeably across both standardized and classroom contexts. As He et al. (2025) caution, such terminological drift can hinder scoping reviews by concealing relevant studies behind unpredictable phrasing. We therefore anticipated that designing a search strategy balancing sensitivity with specificity would prove challenging. Our broad search string comprised ("test*" OR "assess*" OR "exam*") AND ("cogniti*" OR "process*" OR "strateg*" OR "skill*"). This approach prioritized sensitivity to achieve comprehensive mapping goals, though we expected a demanding screening phase that might entail adaptation. This exemplifies the core scoping review challenge: defining and capturing the intended scope effectively while acknowledging that iterative refinement occurs as familiarity with the literature grows (Arksey & O'Malley, 2005).

Defining inclusion boundaries presented another anticipated challenge: How comprehensive is enough? This question guided our inclusion criteria: (a) focusing on L2 learners' cognitive processes during language-test tasks, excluding research on learning or instruction; (b) establishing a 2005–2023 timeframe anchored by Weir's socio-cognitive framework; (c) restricting the corpus to English-language studies; and (d) including journal articles, books, book chapters, research reports, and doctoral theses, while excluding master's dissertations due to retrieval challenges. Each decision represented a conscious trade-off between comprehensiveness and practicality. Including doctoral theses aimed to capture potentially crucial unpublished work, aligning with Arksey & O'Malley's (2005) call for broad searches. Conversely, excluding non-English studies and setting temporal boundaries limited scope and potentially introduced bias due to feasibility constraints. The challenge lay in achieving a justifiable balance with clearly defined and transparently reported boundaries (Peters et al., 2020).

We anticipated that applying inclusion criteria consistently across thousands of records would challenge multiple reviewers, particularly given the conceptual ambiguity across all language skills. The risk of subjective drift and disagreement was substantial. Similarly, charting diverse data from varied study types such as qualitative reports, quantitative experiments, mixed-method research, and technical descriptions into a standardized form risked inconsistent interpretation. To enhance reliability and transparency, we implemented several strategies following established best practices (Levac et al., 2010; Peters et al., 2020): developing explicit criteria, piloting the charting form, conducting partial dual review for calibration, utilizing shared Excel spreadsheets for tracking, and maintaining a detailed researcher logbook (Liu & Chong, 2023).

We also expected significant methodological diversity among included studies, encompassing varied designs, methods ranging from think-alouds to eye-tracking, and diverse study populations. This heterogeneity would inevitably complicate synthesis. Following Arksey and O'Malley (2005) and Peters et al. (2020), we planned to use thematic analysis (Nowell et al., 2017) to descriptively map the range of approaches and findings rather than attempt an aggregative summary, which would be inappropriate given the expected diversity. The challenge lay in developing a nuanced yet coherent synthesis from disparate elements, a recognized characteristic of scoping review outputs.

In addressing these foreseeable challenges, we assembled a research team balancing substantive and methodological expertise: two senior researchers and two early-career researchers, deliberately pairing language-testing specialists with research-synthesis experts. This combination promised the "powerful synergy" highlighted by Chong and Plonsky (2024, p. 1517), though we recognized that researchers without a language testing background would require additional support for screening and extraction tasks. As Section 2.2 details, this anticipated challenge of expertise differentials manifested more profoundly than expected, particularly in distinguishing between cognitive processes in testing versus learning contexts.

Beyond these specific challenges, we recognized a fundamental principle: the protocol must function as a living document rather than a rigid prescription. This understanding, emphasized throughout the scoping review literature (Arksey & O'Malley, 2005) and demonstrated by researchers like Terrier et al. (2025), reflects what Chong (2025) describes as the "double duty" of synthesis guidelines (p. 7). This double duty requires balancing standardization for methodological rigor with flexibility to accommodate the inherent heterogeneity of research contexts and evidence types. Our published review protocol (Chan & Chong, 2023) embodied this dual nature, providing systematic structure while acknowledging that engaging with the literature would inevitably reveal unanticipated complex-

ities requiring justified modifications. The key was maintaining transparency about all changes while preserving the methodological integrity that distinguishes scoping reviews from traditional literature reviews. As the following section illustrates, this principle would be tested repeatedly when we encountered challenges that exceeded even our careful preparations.

2.2 Emergent dilemmas: Confronting unexpected realities

The dynamic process of conducting the review surfaced three additional challenges requiring adaptive responses. Each tested our ability to maintain systematic rigor while responding to unexpected realities.

The first major challenge involved identifying appropriate information sources. Although our published review protocol listed seven databases, two proved non-functional: neither the Directory of Open Access Journals (DOAJ) nor the Electronic Thesis Online Service (EThOS) yielded any relevant studies, despite iterative searches by two team members using our finalized search strings. This necessitated focusing on the remaining five databases: EBSCO (including ERIC), ProQuest, Scopus, PsycINFO, and SSCI. To ensure comprehensive coverage, we supplemented our database search with manual searches of research reports from major language test providers (IELTS, TOEFL, Pearson, Aptis, and GEPT). While this approach yielded 1,735 studies (1,699 from databases, 36 from manual searches), our initial corpus analysis revealed a critical gap: several highly relevant studies by a key researcher were missing. This exposed the inherent limitations of keyword-based searches in capturing the full breadth of relevant literature. As Terrier et al. (2025) observe, keyword searches "can narrow the scope by relying on precise search terms, potentially excluding studies that use alternative phrasing linked to a specific theoretical framework" (p. 3). These insights prompted us to develop a comprehensive remedial strategy combining manual journal searches, citation tracing, and network analysis tools, detailed in Section 3.

The second challenge involved screening and charting the identified studies. While our 1,735 records represented a moderate volume compared to many scoping reviews, the nature of our screening requirements transformed this manageable number into a significant challenge. Two interrelated factors (the conceptual ambiguity in our review topic and the domain-specific expertise it requires to screen candidate studies) necessitated intensive full-text examination of nearly every study, dramatically increasing the screening workload beyond what the raw numbers suggested. First, the conceptual ambiguity proved more profound than anticipated. Many studies employed relevant keywords in their titles or abstracts but lacked actual process data meeting our inclusion criteria. For instance,

numerous papers discussed cognitive processes theoretically without providing empirical investigation of these processes during test performance. Determining eligibility therefore required meticulous examination of methodology sections to ascertain whether studies empirically investigated cognitive processes during language testing rather than learning contexts. Abstract screening alone proved insufficient. Second, this conceptual complexity was compounded by the expertise requirements. One team member without specialized language testing background experienced persistent difficulty distinguishing between studies examining cognitive processes for learning purposes versus those focused on test performance. This revealed that our seemingly straightforward inclusion criteria actually required nuanced understanding of language testing contexts. While we had anticipated that team members without language testing expertise would need additional support (Section 2.1), the extent exceeded our expectations. The combination of conceptual ambiguity requiring full-text examination and the need for specialized expertise meant that our 1,735 studies demanded far more intensive screening than initially projected. Consequently, we shifted from full to partial dual screening: the two team members responsible for screening initially reviewed approximately 10% of studies together to calibrate their judgments, then independently screened the remaining studies from their assigned databases. This experience illustrated how literature volume, conceptual complexity, and team expertise interact to shape scoping review feasibility, forcing us to operationally define "comprehensive enough" within our resource constraints.

The third emergent challenge arose from our exploratory trial of AI tools for data extraction. Motivated by the anticipated workload of charting numerous included studies, we tested two Large Language Models (LLMs) using ChatGPT Pro (GPT-4) and Google Gemini (2.0) chat interfaces to potentially enhance extraction efficiency. A set of randomly selected studies across publication types were submitted to both platforms for data extraction, with the results being evaluated by the whole review team. Our trial proved partially successful. First, the LLMs competently extracted factual bibliographic information, but they struggled with tasks requiring interpretation. We found that the AI consistently misidentified process-tracing methodologies, conflated different cognitive assessment tools, and could not distinguish between process and product measures. These distinctions, however, were crucial for our analysis. Second, technical constraints proved equally challenging. Both tools' limited context windows necessitated splitting longer papers (particularly doctoral theses) across multiple chat sessions, requiring repeated retraining of the model and likely contributing to extraction inconsistencies. Each new chat window essentially reset the tool's understanding of our extraction requirements, compounding accuracy issues. These limitations, both conceptual and technical, produced error levels requiring such extensive

human verification that we reverted to fully manual extraction. However, this experience provided valuable insights into distinguishing between mechanical and interpretive tasks in research synthesis. While AI failed at interpretive categorization requiring domain knowledge, it showed promise for structured, mechanical tasks. This reflects what Wagner et al. (2026, p. 13) term a 'varying degrees of augmentation' scenario, in which GenAI effectively supports mechanical tasks while human expertise remains indispensable for interpretive synthesis. A growing body of evidence corroborates this distinction: Clark et al. (2025) reported incorrect data extractions ranging from 4% to 31% across studies in a systematic review of GenAI use in evidence synthesis; Lieberum et al. (2025) concluded that fully validated LLM applications for systematic reviews remain lacking; and within Applied Linguistics, Hampson et al. (2025) found that GenAI achieved acceptable reliability for screening and methodology identification but failed at tasks requiring implicit reasoning. Notably, recent developments — including the expansion of context windows to one million tokens (e.g., Gemini, Claude) and the emergence of agentic AI tools capable of working directly with local file systems (e.g., Claude Cowork) — suggest that many of the technical barriers we encountered are rapidly diminishing. These advances, combined with our understanding of AI's current capabilities and limitations, inform our reflections on a potential human-led, AI-facilitated approach to research synthesis (Section 4).

Beyond their methodological implications, several of the challenges described above also carry ethical dimensions, from the representativeness concerns raised by omitting stakeholder consultation to the trustworthiness implications of adapting screening procedures. Among these, the most salient considerations emerged from our exploratory use of commercial AI platforms for data extraction. As Hampson et al. (2025) note, uploading research articles to commercial platforms raises questions about data transfer to third-party servers, a concern that recent studies have addressed by disabling model training on uploaded content to protect the intellectual property of original authors (Aryadoust & Jia, 2026). More broadly, the normalisation of AI-assisted review processes without rigorous validation risks deskilling researchers and undermining the critical engagement that characterises rigorous synthesis (Hampson et al., 2025), underscoring the need for emerging ethical frameworks to establish standards for responsible AI integration in evidence synthesis (see Harasgama et al., 2026, for a comprehensive overview). These considerations lend retrospective support to our decision to revert to fully manual extraction. However, whether fully manual processes will remain the standard for research synthesis as AI capabilities continue to evolve remains an open question.

These emergent dilemmas underscore the importance of maintaining flexibility and adaptability throughout the review process. They reinforced the need

for robust team communication channels and highlighted the value of transparently documenting all adaptations to the original protocol. Most importantly, these challenges exemplified the iterative and reflective nature of the scoping review process advocated by Levac et al. (2010), demonstrating that even well-planned reviews must remain responsive to the realities encountered during implementation.

3. Course of action

The challenges encountered, both anticipated and emergent, informed our operational approach to conducting the scoping review. Our implementation involved systematic progression through four key stages: search strategy implementation, screening, data charting, and synthesis. While we largely adhered to our published protocol (Chan & Chong, 2023), the realities of execution necessitated strategic adaptations. These modifications responded particularly to the incomplete database coverage discovered during initial analysis and insights gained from our AI exploration trial. This section details both our implemented procedures (Sections 3.2–3.3) and our current work, including planned supplementary searches (Section 3.1) and ongoing synthesis activities (Section 3.4).

3.1 Supplementary search strategies

Following our discovery of missing key studies during initial corpus analysis, we implemented a multi-pronged supplementary search strategy to ensure comprehensive coverage before proceeding with synthesis. This adaptive response exemplifies the iterative refinement that characterizes rigorous scoping reviews (Levac et al., 2010).

First, we conducted systematic hand-searches of key journals in language testing, specifically *Language Testing* and *Language Assessment Quarterly*, covering the entire 2005–2023 timeframe. These journals were selected based on their central role in publishing cognitive process research in language assessment. Second, we employed bidirectional citation tracing of seminal studies identified in our initial corpus, examining both references cited and subsequent citations. Additionally, we searched for the professional profiles of key researchers identified through our initial analysis across Google Scholar, ResearchGate, and institutional websites. Third, we leveraged AI-powered network analysis tools including Connected Papers (https://www.connectedpapers.com/) and Research Rabbit (https://researchrabbitapp.com/) to map citation networks and identify studies that keyword searches may have missed. This targeted application of AI for bib-

liographic discovery, in contrast to our only partially successful attempt at AI-assisted data extraction, illustrates the importance of matching technological tools to appropriate tasks. As Figure 2 demonstrates, using Chan (2017) as a seed paper in Connected Papers generated a network visualization revealing interconnected studies, enabling systematic verification of our database coverage.

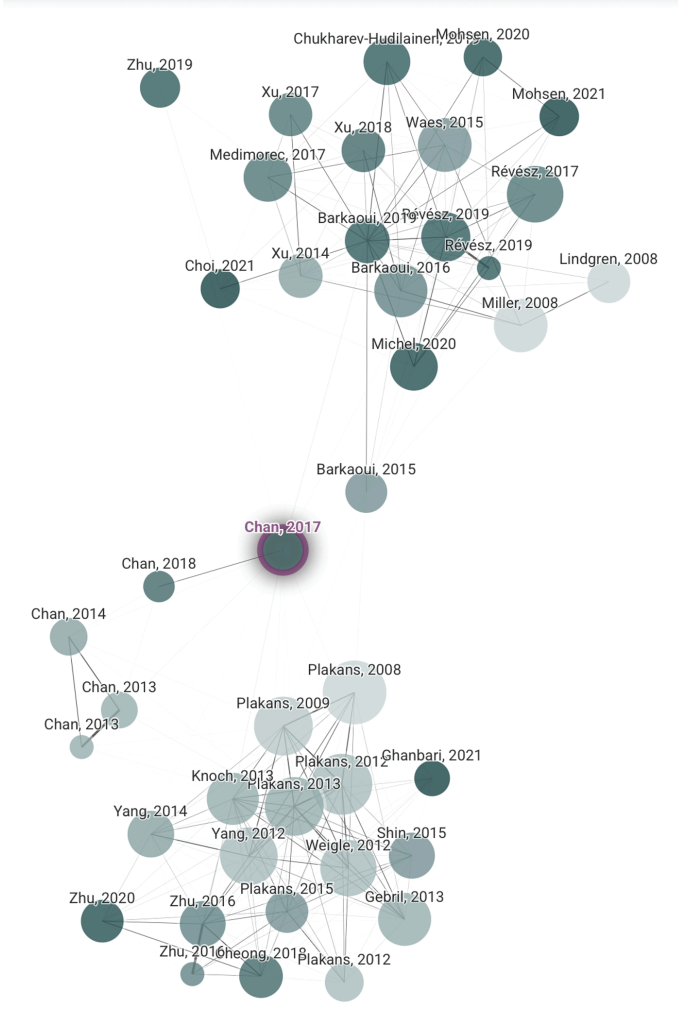


Figure 2. Citation network analysis of Chan (2017) via Connected Papers

These supplementary searches directly addressed the limitations of keyword-based searches identified by Terrier et al. (2025), who found that such searches "can narrow the scope by relying on precise search terms, potentially excluding

studies that use alternative phrasing" (p. 3). Completion of these searches ensured comprehensive coverage before we proceed to final analysis and synthesis. Notably, the effectiveness of Connected Papers and Research Rabbit for comprehensive bibliographic discovery has gained empirical validation across multiple disciplines, with institutional guidance now providing detailed implementation protocols (Birmingham City University Library & Learning Resources, 2025).

3.2 Adaptive screening protocol

To manage the substantial volume while maintaining rigor, we implemented an adapted screening protocol that balanced feasibility with systematic review standards (Arksey & O'Malley, 2005; Peters et al., 2020). Our approach evolved through iterative refinement based on team expertise and emerging challenges.

For title and abstract screening, we began with calibration exercises where two team members screened and compared results on 5% sample sets. Following this calibration, we proceeded with a streamlined single-reviewer approach, with each team member assessing approximately 50% of records. Reviewers applied inclusion criteria conservatively, excluding only clearly irrelevant studies to minimize false negatives at this preliminary stage.

Full-text screening demanded more intensive collaboration due to the conceptual ambiguity distinguishing cognitive processes in testing versus learning contexts. We implemented a systematic calibration protocol beginning with two norming sessions led by our language testing expert to establish shared understanding of inclusion criteria. During these sessions, a random 5% sample underwent independent dual review by two designated screeners strategically paired: one with research synthesis expertise, the other with language testing expertise. Both completed standardized forms requiring explicit justification of decisions, specifically addressing whether studies were assessment-oriented and examined cognitive processes. Following initial reviews, full-team calibration meetings resolved discrepancies through discussion. These discussions crystallized into a guiding question for screeners: *Does the study focus on/shed insights into cognitive processes that test takers engage in during reading/writing/speaking/listening tests for summative and/or formative purposes?* We supplemented this with two exemplary studies (Révész et al., 2017; Yu et al., 2017) as reference anchors and encouraged flagging uncertain cases for team consultation.

While this hybrid approach departed from full dual-review gold standards (Higgins & Green, 2011), it aligned with Levac et al.'s (2010) emphasis on pragmatic teamwork and reflects common practice in large-scale reviews where resource constraints necessitate adaptation (Tricco et al., 2016). All decisions were systematically logged in shared Excel spreadsheets with standardized exclusion

codes (e.g., "wrong population," "no process data," "theoretical only") and cross-referenced in our researcher logbook, ensuring complete transparency for PRISMA reporting (Page et al., 2021).

3.3 Manual data charting implementation

For studies meeting final inclusion criteria, we implemented data charting using a structured form adapted from Chong and Reinders (2022). The form underwent rigorous piloting through a two-stage process. First, all four team members independently reviewed 5–10 papers each to establish common understanding. Subsequently, two researchers independently charted a diverse subset representing approximately 20% of included studies. Through systematic comparison and discussion, we refined the form to ensure item clarity, confirm its capacity to capture diverse data types, and establish consistent charting procedures, aligning with established rigor recommendations (Levac et al., 2010; Peters et al., 2020). The finalized form captured five key categories: publication details, research design and participants, proficiency and assessment information, methodological approaches, and findings.

Following our partially successful AI trial, we committed to full manual extraction. The intensive workload this decision entailed necessitated recruiting an additional team member who was a doctoral researcher specializing in cognitive processing research. This expansion of our team exemplified the adaptive resource management often required in comprehensive reviews.

We implemented a systematic training protocol for the new member. They independently examined and coded the 24 studies that were used for the AI data extraction trial, thoroughly reading each paper and entering relevant information into our pre-defined Excel template. Our experienced language testing researcher then verified these entries, with structured meetings addressing uncertainties and establishing coding consistency. Following successful training completion, the new member proceeded with charting the remaining studies, while the experienced researcher conducted periodic quality checks to maintain coding reliability throughout the process.

This manual approach, while time- and labor-intensive, ensured the conceptual accuracy necessary for meaningful analysis. Such protocol adaptations, including team composition adjustments, reflect the realities of conducting rigorous scoping reviews and align with broader recognition that methodological flexibility serves rather than compromises systematic inquiry (Chong, 2025; Tricco et al., 2018).

3.4 From charted data to integrated findings

While our synthesis remains in progress, outlining our analytical approach completes the methodological narrative and illustrates how scoping reviews move from data collection to meaningful insights.

Our synthesis employs thematic analysis following Nowell et al.'s (2017) six-phase framework for ensuring trustworthiness: familiarization through repeated reading; systematic initial coding; searching for themes aligned with research questions; iterative theme refinement through team discussion; defining and naming final themes; and producing the narrative report. This qualitative synthesis is complemented by descriptive numerical analysis of key variables (publication trends, skill distribution, methodological approaches) directly coded in our Excel database.

The final report will integrate quantitative patterns and qualitative themes to address our research questions, explicitly identifying gaps (e.g., under-researched skills) and methodological trends (e.g., emerging process-tracing tools). Following PRISMA-ScR guidelines (Tricco et al., 2018), we will transparently acknowledge limitations including our English-only scope. This systematic yet flexible approach exemplifies how scoping reviews can map complex, evolving research domains while maintaining methodological rigor. The synthesis findings will be presented through both narrative and visual formats, with figures designed to enhance accessibility and comprehension of complex patterns.

4. Alternative course of action

This section analyzes three alternative pathways that could have substantially shaped our review, offering insights for future researchers navigating similar methodological decisions.

The choice between partial and full dual screening represents a fundamental trade-off in scoping review methodology. While independent dual review of all 1,735 records at both screening stages would have maximized protection against error and bias (Higgins & Green, 2011), this gold standard proved unfeasible within our resource constraints, a reality commonly forcing pragmatic adaptations in large-scale reviews (Tricco et al., 2016). Our partial screening approach carried two primary risks. First, despite calibration efforts, different reviewers might apply inclusion criteria inconsistently, introducing subjective variation. Second, single-reviewer screening increases the possibility of overlooking eligible studies, particularly those with ambiguous titles or abstracts. These risks are not merely theoretical, as the potential for inconsistent application and missed studies

increases with single-reviewer screening. To mitigate these risks, we implemented three key safeguards: conservative initial screening that favored inclusion when uncertain, iterative calibration sessions using shared samples to align reviewer judgments, and systematic escalation of ambiguous cases for team consensus. Our 5% dual-screening samples served as ongoing quality checks, revealing acceptable agreement levels that justified our approach. This methodological choice exemplifies the tension between ideal standards and practical constraints that defines much of scoping review practice. Rather than viewing such adaptations as compromises, we frame them as conscious decisions that balance rigor with feasibility, provided they are transparently reported and systematically implemented (Chong, 2025). Future reviews facing similar constraints might consider our hybrid model as one viable approach among feasible options.

Stakeholder consultation is recognized as a valuable optional sixth stage by Arksey and O'Malley (2005); however, it was ultimately omitted from our review process. This decision represents another significant methodological trade-off. Engaging diverse stakeholders could have enriched our review in multiple ways. Test developers, language testing researchers, Second Language Acquisition (SLA) experts, teachers, and L2 learners might have provided crucial insights into how different communities understand and differentiate cognitive strategies from skills, potentially helping resolve the terminological ambiguity we encountered. Given the interdisciplinary nature of language testing and its historical integration of SLA research findings, consultation with researchers from adjacent fields could have offered theoretical frameworks to guide our synthesis. Moreover, as multimodal assessment tasks increasingly transform the field, stakeholder input could have clarified how current understanding of cognitive processes should inform task design and validation. The practical benefits would have been substantial. Stakeholder consultation enhances both relevance and applicability, ensuring that mapped evidence addresses real-world concerns (Oliver, 2001). This seems particularly critical given our finding that key researchers' work was initially missed by database searches and stakeholder input might have identified such gaps earlier. However, meaningful stakeholder engagement demands considerable resources: time for recruitment and consultation, personnel for coordination, and funding for potential compensation. These requirements led to our pragmatic decision to omit this stage, a choice frequently observed in published scoping reviews, suggesting similar constraints affect many research teams (Tricco et al., 2016). This omission likely resulted in findings that, while methodologically rigorous, may be less immediately applicable to practice settings and less responsive to stakeholder priorities. This trade-off between research efficiency and translational impact remains a persistent challenge in scoping review methodology.

A third significant methodological decision involved whether and how to integrate artificial intelligence tools into our data extraction and synthesis processes. While our decision to abandon AI integration was primarily driven by technological limitations at the time, reflection on this experience alongside recent advances (Susnjak et al., 2025) suggests that a 'human-led, AI-facilitated' framework could offer a more nuanced approach. This principle, developed through our direct experience, aligns with emerging empirical evidence on AI capabilities across the research synthesis workflow. Within Applied Linguistics, this evidence is now building. Hampson et al. (2025) found that GenAI achieved acceptable reliability for mechanical tasks such as screening and methodology identification but failed at tasks requiring implicit reasoning, while Aryadoust and Jia (2026) successfully employed AI as a second coder for structured annotation in a review of 237 L2 studies, achieving near-perfect inter-coder reliability. Both studies reinforce the mechanical-interpretive boundary that our own experience identified.

Our experience revealed a crucial distinction between mechanical and interpretive tasks that could guide strategic AI deployment across the scoping review process. This principle would match AI capabilities to appropriate tasks while maintaining human oversight over all critical decisions. For study identification (Stage 2), AI tools could effectively support mechanical tasks such as duplicate detection and initial document retrieval. Our successful use of Connected Papers and Research Rabbit for citation network analysis exemplifies how AI can enhance coverage by identifying studies missed through traditional searches. During screening (Stage 3), AI could assist with initial title/abstract filtering, which is largely a mechanical process. However, researchers must retain full control over inclusion decisions requiring theoretical judgment. The mechanical-interpretive distinction becomes most pronounced during data extraction (Stage 4). AI demonstrates reliability for extracting factual elements: publication details, sample sizes, and basic study characteristics. However, our trial confirmed that conceptual categorization remains beyond current AI capabilities. This includes distinguishing the process from product measures and identifying cognitive assessment types. Recognizing this boundary earlier would have prevented wasted effort on tasks beyond AI's current capabilities and led us to a more balanced understanding of AI's appropriate role in research synthesis. For synthesis (Stage 5), while AI might assist with data organization and visualization, thematic analysis and interpretation must remain entirely human-led. Current AI tools lack the conceptual understanding required for pattern identification and meaning-making that characterises rigorous synthesis (Clark et al., 2025; Lieberum et al., 2025; Susnjak et al., 2025), with Peters and Chin-Yee (2025) further demonstrating

that these tools systematically overgeneralise scientific conclusions even when explicitly prompted for accuracy.

Implementing such a principle positions researchers in multiple essential roles. Researchers must be equipped with AI literacy to understand both the capabilities and limitations of these tools, enabling them to establish validation protocols that prevent over-reliance on automation and ensure transparent documentation (Chong, 2025) of all AI-assisted processes. Simultaneously, they must maintain the analytical engagement necessary for developing synthesis expertise (McKinley et al., 2025). Our proposed approach positions AI as methodological scaffolding rather than replacement, aligning with calls for responsible integration that preserves human judgment in critical decisions while leveraging computational efficiency where appropriate (Harasgama et al., 2026; Mozelius & Humble, 2024). Nie et al. (Chapter 18, this volume) reach a similar conclusion from their reflection on a large-scale meta-review, finding that while AI tools enhanced efficiency within a structured methodological framework, human expertise remained essential for interpretation, resolving uncertainty, and ensuring conceptual coherence.

Upon reflection, the considerable learning curve and validation requirements associated with implementing such a framework, combined with technological limitations at the time of our review, explain why our AI integration attempt proved only partially successful. Nevertheless, this experience yields valuable insights for future researchers. A human-led, AI-facilitated approach that strategically deploys AI for mechanical tasks while preserving human authority over interpretive work could enhance efficiency without compromising rigor. Success requires not only technological readiness but also researcher AI literacy, robust validation protocols, and unwavering commitment to transparency. As AI capabilities rapidly evolve, our experience offers a principled path for integration that upholds the methodological transparency and systematic rigor that define quality scoping reviews.

5. Conclusion and lessons learned

Our intensive engagement with this scoping review on L2 cognitive processes in language testing proved a rich learning experience, yielding six critical insights: (1) protocols must balance systematic structure with flexibility for justified adaptations; (2) practical constraints necessitate transparent compromises, exemplified by our shift from full to partial dual screening; (3) synthesizing heterogeneous literature requires sophisticated analytical mapping beyond simple description, supported by tools like structured Excel charting and systematic the-

matic analysis; (4) collaborative team processes emerge as essential, with systematic approaches demanding coordinated expertise that no single researcher can provide, a reality reinforced by broader trends showing systematic reviews increasingly conducted by teams rather than individuals (Khany et al., 2025); (5) meticulous documentation through detailed logbooks and reporting standards ensures trustworthiness; and (6) while AI tools show promise for mechanical tasks like citation network analysis, our trial confirmed current limitations for conceptual data extraction, reinforcing that human expertise remains irreplaceable for interpretive work.

These lessons underscore that comprehensiveness in scoping reviews is necessarily bounded by practicality, reflecting the broader evolution in language testing toward more systematic yet resource-intensive methodologies. The goal therefore becomes achieving maximum possible coverage within defined constraints through systematic, transparent methods. Achieving this goal demands distinct competencies: strategic scope definition, qualitative synthesis skills, cross-expertise team coordination, adaptive problem-solving, and increasingly, AI literacy to discern appropriate applications of emerging technologies. This shift toward systematic approaches not only enhances methodological rigor but also correlates with greater scholarly impact, suggesting that the additional resources required yield meaningful returns. The field-specific SMART tool (Chong, 2025), based on principles of Systematicity, Transparency, Accessibility, and Reflexivity, offers valuable scaffolding for navigating these methodological complexities.

Despite challenges, this methodology proved effective for mapping our complex research domain. The comprehensive map generated promises to identify gaps, inform future research, and bridge Weir's socio-cognitive framework with emerging multimodal assessment practices.

6. Questions for discussion

1. How should scoping review teams operationalize and justify their "comprehensiveness" when designing search strategies, given the inherent tension between exhaustive coverage and practical feasibility?
2. What calibration methods can ensure consistent application of inclusion criteria when screening literature in fields with significant terminological ambiguity, such as cognitive process research?
3. How should researchers transparently report methodological adaptations (e.g., partial dual screening, AI tool trials) while maintaining credibility and acknowledging limitations?

4. Should stakeholder consultation be reconsidered as essential rather than optional in interdisciplinary fields where conceptual ambiguity and rapidly evolving practices pose significant challenges?
5. What frameworks should guide the integration of AI tools in research synthesis, particularly for distinguishing between mechanical tasks suitable for automation and interpretive tasks requiring human expertise?

Further reading

Arksey, H., & O'Malley, L. (2005). Scoping studies: Towards a methodological framework. *International Journal of Social Research Methodology*, *8*(1), 19–32.

This foundational paper establishes the six-stage methodological framework that serves as a practical guide for scoping review studies.

Chong, S.W. (2025). Synthesis Methods and Reporting Tool (SMART) for research syntheses in Applied Linguistics. *Research Synthesis in Applied Linguistics*, *1*(1), 17–38.

This paper introduces the field-specific Synthesis Methods and Reporting Tool (SMART), which assists researchers in ensuring the Systematicity, Transparency, Accessibility, and Reflexivity (STAR) of a review.

Chong, S.W., & Plonsky, L. (2024). A typology of secondary research in Applied Linguistics. *Applied Linguistics Review*, *15*(4), 1569–1594.

This paper outlines a comprehensive framework that enables researchers to distinguish among 13 types of reviews in Applied Linguistics, thereby facilitating the selection of the most appropriate synthesis method for their research questions.

Terrier, L., Garnier, M., & Ali, S. (2025). Leading a scoping review on L2 pronunciation: Some key elements of methodology. *Research Methods in Applied Linguistics*, *4*(3), 1–13.

This paper provides an exemplary methodological reflection on conducting a scoping review in Applied Linguistics.

Susnjak, T., Hwang, P., Reyes, N., Barczak, A.L.C., McIntosh, T., & Ranathunga, S. (2025). Automating research synthesis with domain-specific large language model fine-tuning. *ACM Transactions on Knowledge Discovery from Data*, *19*(3), Article 68.

This paper proposes a framework for AI-facilitated research synthesis that offers insights into responsible technology integration while acknowledging current limitations.

AI use declaration statement

During the preparation of this work, the author used Napkin.ai for creating visual representations of the review process (Figure 1) and Claude (Opus 4.6) for proofreading and reference verification.

As author, I confirm that after using the tool(s), I reviewed and edited the content as needed and take full responsibility for the content of the publication, in alignment with COPE and the publisher's AI policy.

References

Aryadoust, V., & Jia, Y. (2026). Univariate normality checking practices in L2 research: An AI-assisted systematic review. *Studies in Second Language Acquisition*, 1–36.

Aryadoust, V., & Luo, L. (2023). The typology of second language listening constructs: A systematic review. *Language Testing*, *40*(2), 375–409.

Birmingham City University Library & Learning Resources. (2025). *Artificial intelligence (AI) and the literature review process*. Library & Learning Resources. Retrieved 22 August 2025 from https://libguides.bcu.ac.uk/generativeAI

Brunfaut, T., & McCray, G. (2015). *Looking into test-takers' cognitive processes while completing reading tasks: A mixed-method eye-tracking and stimulated recall study*. British Council. https://www.britishcouncil.org/sites/default/files/brunfaut_and_mccray_report_final_o.pdf

Chan, S. (2017). Using keystroke logging to understand writers' processes on a reading-into-writing test. *Language Testing in Asia*, *7*, Article 10.

Chan, S., & Chong, S.W. (2023). Protocol for a scoping review of L2 learners' cognitive processes in language testing (IDESR000086). *International Database of Education Systematic Reviews*. https://research-repository.st-andrews.ac.uk/bitstream/handle/10023/28736/IDESR000086_14_07_2023.pdf?sequence=1

Chong, S.W., & Reinders, H. (2022). Autonomy of English language learners: A scoping review of research and practice. *Language Teaching Research*, *29*(2), 607–632.

Clark, J., Barton, B., Albarqouni, L., Byambasuren, O., Jowsey, T., Keogh, J., Liang, T., Moro, C., O'Neill, H., & Jones, M. (2025). Generative artificial intelligence use in evidence synthesis: A systematic review. *Research Synthesis Methods*, *16*(4), 601–619.

Hampson, T., Cargos, K., & McKinley, J. (2025). Automate the 'boring bits': An assessment of AI-assisted systematic review (AIASR). *Research Methods in Applied Linguistics*, *4*(3), 100258.

Harasgama, S., Pearce, H., Loftus, L., Painter, H., & Ford, J. (2026). An overview of artificial intelligence approaches for automating evidence synthesis. *Public Health*, *254*, 106220.

He, S., Sénécal, A.-M., Stansfield, L., & Suvorov, R. (2025). A scoping review of research on second language test preparation. *Language Testing*, *42*(1), 11–47.

Higgins, J.P.T., & Green, S. (Eds.). (2011). *Cochrane handbook for systematic reviews of interventions* (Version 5.1.0). The Cochrane Collaboration. https://www.handbook.cochrane.org

Khany, R., Fahimpour, Z., & Teo, T. (2025). Exploring secondary research studies in Applied Linguistics: A scoping review of methodologies, trends, and impact. *Research Synthesis in Applied Linguistics*, *1*(1), 100–129.

Levac, D., Colquhoun, H., & O'Brien, K.K. (2010). Scoping studies: Advancing the methodology. *Implementation Science*, *5*, Article 69.

Lieberum, J.-L., Töws, M., Metzendorf, M.-I., Heilmeyer, F., Siemens, W., Haverkamp, C., Böhringer, D., Meerpohl, J. J., & Eisele-Metzger, A. (2025). Large language models for conducting systematic reviews: On the rise, but not yet ready for use — A scoping review. *Journal of Clinical Epidemiology*, *181*, 111746.

Liu, Q., & Chong, S. W. (2023). Bilingual education in China: A qualitative synthesis of research on models and perceptions. *Applied Linguistics Review*, *15*(4), 1671–1693.

McKinley, J., Chen, X., Sun, Y., Xing, W., Fang, C., & Hampson, T. (2025). Training novice researchers through systematic reviews: Methodological insights from TESOL master's projects. *Research Synthesis in Applied Linguistics*, *1*(1), 39–52.

Mozelius, P., & Humble, N. (2024). On the use of generative AI for literature reviews: An exploration of tools and techniques. *European Conference on Research Methodology for Business and Management Studies*, *23*(1), 161–168.

Munn, Z., Peters, M. D. J., Stern, C., Tufanaru, C., McArthur, A., & Aromataris, E. (2018). Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. *BMC Medical Research Methodology*, *18*(1), 143.

Nowell, L. S., Norris, J. M., White, D. E., & Moules, N. J. (2017). Thematic analysis: Striving to meet the trustworthiness criteria. *International Journal of Qualitative Methods*, *16*(1), 1–13.

Oliver, S. (2001). Making research more useful: Integrating different perspectives and different methods. In S. Oliver & G. Peersman (Eds.), *Using research for effective health promotion*. Open University Press.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *Systematic Reviews*, *10*, 89.

Peters, U., & Chin-Yee, B. (2025). Generalization bias in large language model summarization of scientific research. *Royal Society Open Science*, *12*(4), 241776.

Peters, M. D. J., Godfrey, C., McInerney, P., Munn, Z., Tricco, A. C., & Khalil, H. (2020). Chapter 11: Scoping reviews. In E. Aromataris & Z. Munn (Eds.), *JBI manual for evidence synthesis*. JBI.

Révész, A., Michel, M., & Lee, M. (2017). Investigating IELTS Academic Writing Task 2: Relationships between cognitive writing processes, text quality, and working memory. *IELTS Research Reports*, *3*, 1–48. IDP: IELTS Australia and British Council.

Stratman, J., & Hamp-Lyons, L. (1994). Reactivity in concurrent think-aloud protocols: Issues for research. In P. Smagorinsky (Ed.), *Speaking about writing: Reflections on research methodology* (pp. 89–112). Sage.

Tricco, A. C., Lillie, E., Zarin, W., O'Brien, K. K., Colquhoun, H., Levac, D., Moher, D., Peters, M. D. J., Horsley, T., Weeks, L., Hempel, S., Akl, E. A., Chang, C., McGowan, J., Stewart, L., Hartling, L., Aldcroft, A., Wilson, M. G., Garritty, C., Lewin, S., Godfrey, C. M., Macdonald, M. T., Langlois, E. V., Soares-Weiser, K., Moriarty, J., Clifford, T., Tunçalp, Ö., & Straus, S. E. (2016). A scoping review on the conduct and reporting of scoping reviews. *BMC Medical Research Methodology*, *16*, Article 15.

Tricco, A. C., Lillie, E., Zarin, W., O'Brien, K. K., Colquhoun, H., Levac, D., Moher, D., Peters, M. D. J., Horsley, T., Weeks, L., Hempel, S., Akl, E. A., Chang, C., McGowan, J., Stewart, L., Hartling, L., Aldcroft, A., Wilson, M. G., Garritty, C., … Straus, S. E. (2018). PRISMA extension for scoping reviews (PRISMA-ScR): Checklist and explanation. *Annals of Internal Medicine*, *169*(7), 467–473.

Wagner, G., Prester, J., Mousavi, R., Lukyanenko, R., & Paré, G. (2026). Generative artificial intelligence for literature reviews. *Journal of Information Technology*.

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Palgrave Macmillan.

Yu, G., He, L., & Isaacs, T. (2017). *The cognitive processes of taking IELTS Academic Writing Task 1: An eye-tracking study*. British Council, IDP: IELTS Australia and Cambridge English Language Assessment. https://search.informit.org/doi/pdf/10.3316/informit.840755372872436

Zhai, J., & Aryadoust, V. (2022). The metacognitive and neurocognitive signatures of test methods in academic listening. *Frontiers in Psychology*, *13*, Article 930075.

CHAPTER 12

Mapping research evidence with a scoping review

A critical appraisal

Inma Álvarez & Mara Fuertes Gutiérrez
The Open University

This chapter aims to provide a reflective account on the process of conducting a scoping review. For these purposes, we use as a case study a review undertaken to explore the breadth and depth of team teaching approaches and practices in languages in higher education institutions. We explore how we designed the research study and navigated the unexpected issues that arose when following the five-stage framework developed by Arksey and O'Malley's (2005) for this type of review. We also critically assess this established methodological approach highlighting potential alternatives and offer guidance for researchers interested in conducting scoping reviews in the field of Applied Linguistics.

Keywords: scoping review, Applied Linguistics, research methodology, research design, research synthesis

Background

Introduction

Scoping reviews (ScR) are a type of research synthesis addressing broad research questions to map the extent, range, or nature of available research on a topic. ScR are still rare in the field of Applied Linguistics. As of March 2025, after examining data from Plonsky's (n.d.) and BAAL's Research Synthesis in Applied Linguistics SIG (RSAL, n.d.) databases, we note that they jointly list 30 published ScR (after removing duplicates) in the discipline. According to Chong (2025), in October 2024, both databases recorded a total of 754 research synthesis (RS), with meta-analysis being the most common type of review. Excluding the four ScR published in 2025 or currently in press, ScR would account for 3.4% of the

https://doi.org/10.1075/rmal.17.12alv
Available under the CC BY-NC-ND 4.0 license.© 2026 John Benjamins Publishing Company

total RS published until then. Although it falls outside the scope of this chapter, it would be useful to determine how many ScR are published in languages other than English, to better understand the current interest in this type of RS by researchers globally. Importantly, in these databases, only one review, a meta-analysis (Ripoll & Aguado, 2014), was written in a language different than English — Spanish. However, an exploratory search reveals there are several ScR in Spanish, including some published in high-impact journals (e.g., Tudela-Isanta & Andújar-Molina, 2023).

Figure 1 below plots the publication of ScR in Applied Linguistics, combining data from Plonsky (n.d.), RSAL (n.d.) and the International Database of Education Systematic Reviews (IDESR, n.d.). The graph shows that the oldest reviews date back to 2017, with peaks in published reviews in 2022 and 2024 and already three published in early 2025, suggesting a growing trend in conducting ScR in this field. Publication peaks may be due to adaptations needed in research activities during the Covid-19 pandemic, when face to face empirical studies were challenged by lockdowns and physical distancing, and there was a move to online research methods, including an increased use of secondary data (Ryan et al., 2024).

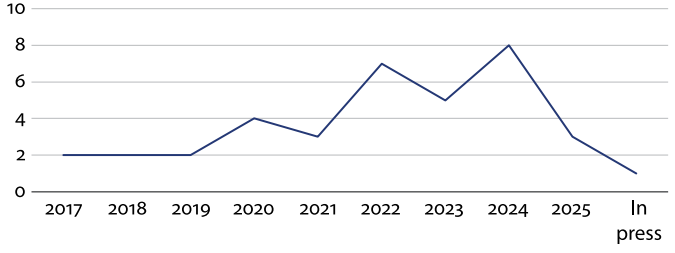


Figure 1. Number of scoping reviews published per year (2017–March 2025)

More specifically, popular topics for ScR in Applied Linguistics include new technologies — in particular gamification (Hung et al., 2018; Zolfaghari et al., 2025) —, assessment (Law, 2024; Rahmati & Sadeghi, 2025), or both (Li et al., 2024), bilingualism (Müller et al., 2020; Feng et al., 2024), and teaching strategies such as flipped classroom (Jiang et al., 2020) or team teaching (Álvarez et al., 2024). Four reviews studied research on English language education (Al-Nafjan & Alhawsawi, 2022; Chong & Reinders, 2022; Li et al., 2024; Zolfaghari et al., 2025) with one interested in world Englishes (Hillman et al., 2021) and one in Arabic as a heritage language (Visonà & Plonsky, 2020). It is also worth noticing that three of these reviews had a focus on teaching English in specific geographical contexts: Saudi Arabia (Al-Nafjan & Alhawsawi, 2022), the Middle East and Africa (Hillman et al., 2021), and Kazakhstan (Satayev et al., 2024). The growing trend

identified together with the breadth of topics covered confirm ScR are increasingly considered a robust method to explore and synthesise existing research in Applied Linguistics, rigorously focusing on the gathering of all the knowledge on a specific topic.

This chapter aims at providing new insights on the process of conducting a ScR in the field of Applied Linguistics to contribute to a better understanding of the conceptual and practical opportunities and challenges in carrying out this type of research synthesis. For these purposes, we draw from our experience in undertaking a ScR on team teaching in Languages Education (Álvarez et al., 2024). We discuss the key issues considered prior to starting the review, the challenges encountered, the strategies employed to address the emerging issues, including an examination of other actions that could have been taken, and the lessons learned throughout the process. This deep reflection on and sharing of our experience conducting a ScR intends to support future researchers interested in doing this type of study.

Research synthesis project

Our research synthesis 'Team teaching in Languages: a scoping review of approaches and practices in higher education', published in the *Innovation in Language Learning and Teaching* journal in 2024, investigated the research evidence available on the extent to which language teaching is conducted collaboratively at universities globally and how it is done. The aim was twofold: on the one hand, we sought to inform the teaching community at universities about the characteristics of the reported experiences, their potential, benefits and drawbacks; on the other hand, we strived to identify any gaps in existing knowledge related to team teaching in languages in higher education to guide future research priorities. Exploring this topic was timely, as it not only can inform collaboration, but it can also support resources optimisation. To conduct this review, we followed the five-stage framework proposed by Arksey and O'Malley (2005), which offers a systematic approach to mapping evidence and ensures transparency and replicability. The findings were evaluated using a recognised typology of team teaching (TT) approaches (Friend et al., 2010). TT practices were identified and reported in detail in a narrative analysis which included implications for language researchers, practitioners and learners in higher education. The review of the research literature revealed that TT practices are used opportunistically and experimentally in classrooms around the world and provided insights on how their main purposes seem to be, on the one hand, teachers' professional development, and on the other, the enhancement of students' learning experience of the study of languages and cultures.

Methodological issues and dilemmas

In this section we present our critical reflections on the methodological issues and dilemmas we faced in our ScR on TT approaches and how we acted upon them. This account illustrates the unexpected setbacks in any research endeavour, including secondary research, as well as the need for resolutive approaches to ensure robustness, accuracy and reproducibility for a methodologically sound study. Table 1 offers an overview of the aspects that we anticipated before conducting the ScR as well as those that emerged while carrying out the study. It also shows how we were committed to resolve them during the process of our research.

Table 1. Methodological issues and dilemmas identified and courses of action

Time	Issues/dilemmas	Courses of action
Anticipated	1. Selection of databases	1. Focused on English texts and Education
	2. Inclusion and exclusion criteria	2. Reconsidered and adjusted criteria
	3. Search strategy	3. Identified sets of relevant key words and spellings
	4. Publication and dissemination	4a. Selected journals with relevant scope
		4b. Selected journals with previous publications of reviews
Emerging	5. Records identified	5. Screened irrelevant data via titles and abstracts
	6. Composition of research team	6a. Introduced additional team member
		6b. Considered user engagement
	7. Use of software	7. Selected compatible reference management software
	8. Data coding	8a. Manual search of abstracts
		8b. Read whole paper to decide suitability

Anticipated issues and dilemmas

The following subsections relate the four issues we anticipated during the design and set up of the research project, and how we resolved them. However, it is important to note that some of these anticipated issues also brought up further dilemmas at different points of the review process so, in a way, it can be said that the issues and dilemmas organically evolved from the start of the project. Therefore, researchers conducting these reviews need to know that while preparation, planning and anticipation at the early stages of the ScR design are essential, it

is almost inevitable that they will need to deal with unexpected issues emerging from both the decisions made but also the data gathered.

1. *Selection of databases*

It is critical that, when compiling research evidence for a ScR, searches are comprehensive and capture relevant and diverse types of literature. For these purposes, the selection of databases to conduct searches for ScR represents a crucial step. Given the large number of databases available, one of the first issues we anticipated was to decide on which databases we should use to ensure an appropriate and thorough collection of the evidence. At the time of our study, no specific guidance of recommended databases for ScR in Applied Linguistics was available, but general advice emphasised the importance of using a range of databases to collect extensive evidence and to reduce the risk of missing key studies (Ewald et al., 2022). No information on optimal combinations of databases to achieve adequate coverage was available either. Furthermore, some databases were not as sophisticated as they are today in terms of their functionalities, and we pre-empted this could cause difficulties in exporting the records selected for screening.

Understanding that the topic of TT was quite niche, we were keen to cast our net wide to capture all the evidence. This is why we initially shortlisted 20 databases, including not only those used more widely in ScR, such as Scopus or Web of Science, but we also consulted others with a specific focus on areas that were relevant to our topic (e.g., ERIC), and those covering specific types of literature (e.g., Proquest Dissertation and Theses). A first step was to test searches in all of them. This is a critical step to check how the database reacts to specific searches using keywords and features that are relevant to the project in hand. Difficulties encountered during the testing ranged from being unable to export data, having to export data individually rather than in batches, or limited string searches. We finally selected 12 major bibliographic electronic databases that capture languages education research in English: Academic Search Complete, CORE, Ebsco Ebooks, Education Abstracts, Education Research Complete, EThOS, NDLTD, Proquest Dissertation and Theses, Science Direct, Scopus, Web of Science, and ZETOC.

Another issue we could pre-empt was related to conducting searches in languages other than English. Our original intention was to conduct the review in the different languages we could read, namely English, French, German, Italian and Spanish. While some databases were supposed to export the literature in a range of languages, in reality, many presented difficulties in exporting records in different languages. After testing searches in the selected languages in several databases with no results, we encounter the ethical dilemma of whether to limit the search to just English as the input language (see Chapter 14, 15 and 17). In the

end, we opted for the practical approach of searching using just research written in English. Taking this decision was also an acknowledgement of the well-known fact that English has become the dominant language of the international scientific community (Ahmed et al., 2023), and therefore the assumption is that the majority, if not all, of the research publications on the selected topic would most likely be found published in this language (see Chapter 10). However, we recognise that this assumption could potentially fail to capture, and therefore represent truly and fairly, teaching practices.

The decision had also a direct impact on our inclusion and exclusion criteria, since the protocol for ScR expects a rigorous declaration of the criteria as part of the study design. We acknowledged in the publication of our ScR the choice of this criterion as an ethical limitation of our study. It is important to note at this point that we have found that linguistic choices in ScR tend to replicate a bias in favour of the literature published in English, which is an issue that affects all disciplines, but it is especially problematic in Applied Linguistics, given the field's focus on areas such as linguistic diversity, language teaching and multilingualism, among others. It must be acknowledged that research is conducted and published in languages other than English, however, these valuable contributions are often overlooked for a variety of reasons (Gironzetti & Fuertes Gutiérrez, forthcoming). In the case of ScR, while traditionally ScR have been limited by the linguistic competence of research teams and the technical limitations of the available databases (as in our case), nowadays it would be advisable to apply a multilingual approach to ScR, using strategies such as the ones we suggest below in the section on "search strategy".

2. *Inclusion and exclusion criteria*

Determining the selection criteria for research evidence to be considered in the ScR, i.e., the characteristics of the studies to be included and excluded, presented another set of dilemmas to consider prior to starting the review. We have already mentioned that the testing of the databases was critical in deciding to include only publications written in English. Furthermore, inclusion and exclusion criteria were refined during the analysis of the evidence, which will be discussed among the emerging issues.

Following the latest recommendations (Peters et al., 2020), we used in our review the Population, Concept, Context Framework to develop our review questions and objectives, defining the characteristics of each element that we wanted to capture in our study. In terms of the target population in our study, participants were language teachers. We encounter the dilemma of deciding whether to include pre-service language teachers. However, being aware that part of their training before becoming qualified teachers frequently includes casual TT, we

anticipated that there would be studies focusing on how this training is conducted and managed within this population. At this point, we decided to exclude interventions with pre-service teachers from our study, since their TT experiences would not reflect an established teaching practice. We focused exclusively on in-service teaching practices which was more closely aligned with our research questions regarding actual classroom activities. The Concept we wanted to research was team teaching pedagogies in higher education. Here, we were presented with the dilemma of including TT practices that just involved languages, that is between language teachers, or the teaching of languages combined with any other disciplines, i.e. language teachers collaborating with teachers of other subjects. Given the importance of practices such as Content and Language Integrated Learning and English as a medium of instruction, we decided to include all TT practices that included at least one language teacher. Finally, the Context was formal teaching approaches in higher education (both at undergraduate and postgraduate level), delivered in online or face to face classrooms in Language Centres or University Departments around the world. This meant that sources of evidence from any country, and in both immersion and non-immersion contexts, were eligible for inclusion. On this occasion, decisions on the Context were straightforward; however, as we will see next, defining this aspect in the search strategy presented some challenges in practice. Overall, these decisions demonstrate the importance of being aware of current practices in languages education to define appropriate inclusion and exclusion criteria.

Two further issues were identified at this stage of the study design. Firstly, an expected part of the inclusion criteria of reviews is the definition of a period of time. This automatically presents a dilemma for the researchers who need to justify their decision regarding the timeframe chosen (see Chapter 6 and 15). Our review was conducted at the beginning of 2020, and the interest was to capture the most recent innovation in TT. For this reason, we focused on studies published during a period of 20 years — from 2000 until the end of 2019. Interestingly, the search did not return any evidence of TT practices before 2005, even though early access papers were included. In this situation, the time period could be adjusted during the testing phase of the search strategy. However, we decided to keep the original period to make a definitive point on the lack of reporting TT practices during the first few years of the 21st century, accentuating the novelty of this approach in the context of higher education.

Finally, another dilemma was the type of evidence that we should take into account. As the intention of the study was to identify and report on empirical evidence, whether quantitative or qualitative, related to TT practices in languages, the decision was to include books, book chapters, journal articles, institutional or pedagogical reports and theses as sources of evidence, following advice for this

type of review. Audiovisual sources and teaching resources related to TT, as well as any studies that were of a theoretical nature, were excluded.

It could be argued that these preliminary decisions are dilemmas that researchers encounter in any type of research synthesis.

3. *Search strategy*

Specificity of the search terms is essential to avoid large quantities of research being found and having to be carefully screened. For this study, a preliminary understanding on the topic and familiarity with the literature related to TT not only helped us identify the questions and objectives of our study, but it also provided very useful knowledge of the keywords we would need for conducting the search (see Chapter 5).

With some initial searches using the core term 'team teaching', we uncovered two other widely used terms to refer to the same type of practice: 'collaborative teaching' or 'coteaching' — sometimes hyphenated as 'team-teaching' and 'co-teaching' —, thus all of them were used in searches. The alternative spelling in most databases did not make any difference in terms of hits, except for NDLTD, where only the hyphenated spelling 'co-teaching' translated into a few results, and Web of Science, where 'co-teaching' showed a larger number of papers ($n = 25$) than the non-hyphenated version ($n = 13$), indicating a preference for such term. We did not find any consistency patterns regarding the spelling preferred by geographical region, but the identification of this variance signalled the importance of identifying the different spellings in English (and potentially in other languages) of the terms selected for the search and the impact this may have for data mining. Therefore, testing how databases react to such subtle differences would add to the rigour of the approach.

Our second search term aimed at gathering the right evidence from the discipline, to avoid getting results irrelevant to languages. After testing, it was clear that the generic term 'language' would help identify studies in modern languages and foreign languages. Similarly, trial searches also included options with reference to the context we were interested in, namely 'university' which could also be referred to as 'higher education' or even 'college' in some educational contexts. Testing of these options revealed that the first two terms were sufficient to identify the relevant literature.

Once the terminological range was captured accurately, we were confident that the relevant data would be found on the chosen databases. Three search strings were combined to be used in searches (Table 2).

Table 2. Main search strings used in the project

1. ("team teaching" or "team-teaching" or "collaborative teaching" or "coteaching" or "co-teaching")
AND
2. ("language")
AND
3. ("university" OR "higher education")

4. *Publication and dissemination*

A final issue we could also anticipate was finding a suitable journal to publish our ScR. Most high-impact journals in the field of Applied Linguistics seem to prioritise empirical research, often favouring studies that present original data collection and quantitative and/or qualitative analysis over secondary research (Dong & Dong, 2024). Therefore, it was crucial to identify a journal with a tradition of publishing secondary research. We pre-selected a set of journals that explicitly welcome in their scope the publication of research syntheses and had indeed samples of these.

Additionally, given the innovative nature of TT in higher education, selecting a journal aligned with pedagogical advancements and emerging trends was important. Since *Innovation in Language Learning and Teaching* actively engages with new approaches to languages education, its scope seemed to align with the aims of our study. Our review not only aimed at contributing to the growing body of ScR in Applied Linguistics but also focused on a developing practice in higher education, showcasing innovation both in research methodology and in its exploration of a novel pedagogical approach to teaching. We would also like to highlight the value of choosing a journal which engages with the OASIS (Open Accessible Summaries in Language Studies) website for the dissemination of summaries "of published journal articles on any language-related topic" (https://www.oasis-database.org/about/), and that offers the opportunity to publish open access.

Emerging issues

Among the challenges encountered whilst conducting this ScR, four of them have been selected for discussion and are presented in chronological order as they emerged.

5. *Records identified*

Despite the careful planning of the search strategy, when screening the titles and abstracts of a total of 5,111 hits, we still found the need to discard most of them, mainly because the practices reported where in contexts outside the scope of the study. For instance, the search captured pre-service teaching practices, which as noted above, were explicitly recorded as an exclusion criterion. In addition, we found papers that focused on teaching at primary and secondary school levels. This reflects a common issue of large databases with insufficient or inconsistent metadata, which can hinder the precision of search queries. Limitations in search algorithms could also result in an overabundance of irrelevant results, particularly in ScR, when keywords used in searches need to be aligned with the broad scope of this type of research synthesis.

6. *Composition of research team*

Regarding the composition of the research team, we would like to refer to two aspects that merit consideration. Firstly, this study was designed by two researchers who also conducted the first steps of the review process, including the selection of the evidence phase. An initial selection of outputs was coded collaboratively to make sure the approach to coding was going to be consistent. A third researcher was added to the team at a later stage for the verification of the data extraction and the analysis of the citations found when it became evident that discrepancies at these later stages of the project could be more easily dealt with if another team member was involved. Research teams of at least three members would be helpful for ScR.

Secondly, recent guidance on ScR suggests knowledge user engagement (Peters et al., 2020). This advice refers to the benefits of a consultation with stakeholders and experts throughout the process, which includes policy makers, practitioners and academics. It should be noted to this respect that the research team for this project ended up consisting of three experienced higher education academics. All of them were familiar the setting of the university classroom and had carried out TT activities in different professional contexts. Therefore, a wider engagement with beneficiaries of the research was not considered necessary at the time.

The unexpected questioning of the composition of the research team led us to an ethical process of self-examination with respect to our identities, bringing to the fore the role played by our potential biases based on previous professional knowledge and experience and prompting us to strive to understand how these were determining our research decisions.

7. *Use of software*

One important early decision was to determine the bibliographic software for exporting large quantities of citations from the databases. We had an institutional license for using Endnote, which is a powerful tool — available online and in a desktop version — where bibliographies can be shared among team members. However, we soon encounter an incompatibility issue due to the *.nbib* extension in some of the documents extracted from the database ERIC, which prevented us from uploading the files onto Endnote. We resolved this issue by selecting a different software, Zotero, which was freely available open-source reference management software and with similar functionality for importing and exporting large quantities of bibliographic data. This software seemed to cope well with the inputs from the different databases.

8. *Data coding*

We encountered several challenges while conducting the coding of the data. Firstly, we applied inclusion and exclusion criteria to screen titles and abstracts, excluding duplicates and several records. However, determining the eligibility of many studies based solely on their title and abstract proved very difficult. Some records lacked abstracts entirely, requiring us to go back to the databases or having to access the papers to locate their abstracts and download them, so we could assess their relevance. Other records lacked sufficient detail in their title or abstracts. The inconsistency in how research titles and abstracts are structured in the discipline was in fact a major obstacle during the selection of relevant studies. Specifically, many failed to indicate the educational context (primary, secondary, higher, etc.) which constituted a key focus of the paper, making it impossible to determine their suitability for our review at this stage. This omission often prevented us from making selection decisions based on this information (title and abstract) alone and having to access the whole paper to make a final decision (see Chapter 6 and 12).

Secondly, even at the stage of full paper reading for appraising the characteristics of the TT interventions and engaging with coding and data extraction, in some cases, we found that some critical details were not identified in the research (see Table 2 in Álvarez et al., 2024). Four out of 14 studies did not specify the number of students in the classroom where team teaching was practiced, and nine did not detailed the frequency and duration of the TT practices analysed in the papers. More challenging was the large amount of missing information regarding the benefits, drawbacks and resulting recommendations of the TT interventions (see Table 3 in Álvarez et al., 2024). This lack of key information affected our capacity to map and synthesize the data from the individual studies to get a larger

perspective on the nature of TT approaches. Obviously, this issue was outside of our control; however it evidently leaves gaps in knowledge creation. Thirdly, another emerging issue was an unexpected lack of methodological rigour found on several of the empirical studies found on this topic, which prevented their replicability. This is, again, an aspect that challenges the quality of the synthesis that can emerge during any type of review of the literature.

These last two challenges related to the coding of the data resonate with issues of the quality of studies in Applied Linguistics identified by Plonsky (2013) in the form of frail methodological approaches and incomplete reporting practices (see Chapter 5 and 17). With regards to this, ScR can be a powerful tool to identify these deficiencies and uncover methodological gaps and weaknesses in the field. Addressing these shortcomings is vital for establishing the discipline on a foundation of robust ethical standards and methodological rigour.

Alternative courses of action

For some of the anticipated and emerging issues we would like to propose some alternative courses of action to support the work of other researchers in considering options for the design and development of their ScR.

Regarding the first issue identified for the selection of databases for the discipline, and in the absence of a consensus on which are the most appropriate, we would like to highlight that perhaps we could have worked more closely with a subject librarian at our institution. Librarians are a good source of advice on the advantages and disadvantages of different databases as well as on adequate approaches to search strings. This could have saved us time experimenting with a large range of databases. This support from librarians could have also helped us decide on an optimal combination of databases for our review. Additionally, consultation with a librarian could have also provided guidance on selecting appropriate software for reference management, potentially avoiding issues encountered with Endnote. To address this, piloting how different software options responded to searches in the databases intended for use might have been beneficial as well.

When possible, searches should be conducted in multiple languages. This makes it necessary to become familiar with non-English databases or to collaborate with scholars from different academic traditions. Such a linguistic approach to reviews shows an ethical commitment of equity, diversity and inclusion of research knowledge and practices (see Chapter 16). Other alternative emerging approaches include using artificial intelligence (AI), which may have a particularly crucial role in translating global research published in a variety of languages.

This is also linked with the emerging issue we faced related to the number of records identified: it would be relevant to test if using AI-driven tools to support the selection of outputs might improve search precision. This technology was not available at the time of our study. Future technology may further facilitate searches and provide researchers with a more accurate set of outputs (see Chapter 12 and 18).

With respect to the anticipated issue on the inclusion and exclusion criteria we established for our study, we realise our approach was quite narrow when applying the PCC framework for Participants. Perhaps considering studies reporting on pre-service teachers participating in TT practices would have resulted in including studies in that context that would have enriched the outcomes of our enquiry.

Looking at the composition of the research team, an alternative course of action in terms of consulting knowledge users before the start of the ScR process and throughout its stages — including protocol development, review process and post review — could involve incorporating one or more language instructors as co-researchers or consultants, favouring a participatory approach to this type of research. In this regard, Bennett and Brunner (2022) indicated the significance of integrating technical, relational and political skills in university-led projects. They also highlighted "the painstaking labour this demands from the participatory researcher" (p. 396), and the critical need for institutional support to navigate this complex approach and the associated tensions, including a series of political and ethical dilemmas that the research team would need to address. Importantly, if the review focuses on higher education pedagogical practices in Applied Linguistics, it is likely that the researchers conducting the review are also practitioners with diverse backgrounds. Therefore, it might not be necessary to include a new member in the team merely for this purpose. Another consideration connected to the research team is that our ScR did not include the calculation of inter-coder reliability. As a result, we found it necessary to seek assistance from a third researcher. Measuring inter-coder reliability would have enhanced the rigour of the process, potentially eliminating the need of an additional researcher later on. A final consideration for future research teams is to consider the inclusion, in the methodological decisions, of an account of their 'positional reflexivity' and/or 'reflexivity of membership' which require a recognition of the point of view of the researched and "a kind of estrangement" with the self (Whitaker & Atkinson, 2021, p. 63). While these reflexive approaches are frequent in ethnographic fieldwork, we believe they have also a place in the methodological considerations of knowledge user engagement in ScR. Explicit anticipatory references to personal and professional self-awareness can ethically guide readers' understanding of how researchers' positionality and membership affect the interpretation of the data.

An alternative course of action to complete the missing data on primary evidence selected for the ScR would have been to approach the authors of these papers to seek additional clarification and fill in the gaps in the information in published studies. This strategy could fall under the recommendation of approaching and involving stakeholders when conducting a ScR. If researchers choose this option, it is crucial to bear in mind that, following recommendations from Aromataris et al. (2024), "a statement should be included of the reviewers' intent to contact authors of primary sources or reviews for further information", and the process needs to be thoroughly documented and explained. Additionally, when reaching out to authors, it is important to proceed with sensitivity, as they might feel scrutinised depending on the nature of the enquiry. Furthermore, if the scope of the ScR includes documenting how the topic is reported in the selected records, establishing direct contact with authors might not be necessary.

A final recommendation is that researchers should consider registering or publishing the details of their ScR protocol, although this is not required. Protocol registration alerts other researchers of the work being done, thus preventing same approaches to reviews taking place simultaneously. The current issue is that ScR are currently not eligible to be published at PROSPERO, one of the most widely-used international systematic review registry, administered by the Centre for Reviews and Dissemination at the University of York (United Kingdom), although they might be accepted in the future, as the website states. Aromataris et al. (2024) suggest registering ScR with the Open Science Framework (https://osf.io/) or Figshare (https://figshare.com/) in the meantime. For Applied Linguistics, another alternative would be to register the protocol in the International Database of Education Systemic Reviews (IDESR), which does cover ScR.

Conclusion

The reflective account presented in this chapter on how we conducted a ScR aiming at expanding our knowledge of the nature of TT practices from the available research has offered us an opportunity to critically examine how we practically and reflexively managed the issues and dilemmas that such a research approach generated. The scrutiny of our own practices conducting this review has also highlighted what alternative courses of action would have improved the quality of our approach, outlining the strengths and weaknesses of the review process.

This opportunity for reflection brought to light how the researcher undertaking a ScR needs to assess and address various dilemmas and issues from the conception of the study. Anticipatory actions will no doubt assist the implementation of the stages of the review; however, researchers will find themselves dealing with

emerging issues not only from the process itself but also from the understanding and interpretation of the data. These dilemmas and issues related to the process of conducting a ScR might be different in nature. On the one hand, researcher might face procedural issues, which are those related to the processes and the steps involved in the review, for instance, the selection of databases or the use of the software. On the other hand, other issues have to do with researchers' conceptual understanding of the actual research, for instance, those arising from the data coding, and therefore could be considered 'epistemological'. Researchers willing to engage in ScR will need to be mindful of the procedural and epistemological tensions emerging from this methodological approach.

Our experience of conducting a ScR also clarified the range of terms used to refer to the same pedagogical approach and identified the real characteristics of such approaches and the factors affecting their practical implementations in the classroom. This conceptual clarification could be seen as one of the potential contributions of these reviews.

Three practical observations for researchers interested in ScR. Firstly, ScR can be lengthy to report in a rigorous manner. Secondly, as it is common in all types of reviews, the majority of the evidence in our ScR was located in international journals, but our study also included among the selected evidence one doctoral thesis, and two publications from conference proceedings. ScR should aim at searching for grey sources of information. Thirdly, Applied Linguistics researchers should lead review practices by critically and systematically considering multilingual approaches in these studies.

Lastly, in terms of recommendations, our ScR helped us identify improvements for research dissemination in Applied Linguistics. Research abstracts should capture the Population, Concept, Context aspects of the studies for clarity and to ensure thoroughness. Publications could be much improved if researchers include specific details of the reported interventions which are sometimes missing in research outputs.

Questions for discussion

1. What specific training do stakeholders (researchers, practitioners, editors, peer-reviewers) need to fully engage with ScR, its reporting requirements and its relevance in the field of Applied Linguistics?
2. What are the challenges and opportunities of involving practitioners throughout the process of conducting a scoping review?
3. How can research in languages other than English (LOTE) be best identified?
4. How can technologies, particularly artificial intelligence, help databases respond more effectively to the search for relevant literature?

5. What are the best strategies to identify relevant grey literature to be included in a ScR?

Further reading

Peters, M. D. J., Godfrey, C., McInerney, P., Munn, Z., Tricco, A. C., & Khalil, H. (2020). Scoping Reviews. In E. Aromataris, C. Lockwood, K. Porritt, B. Pilla, B., & Z. Jordan (Eds.), *JBI manual for evidence synthesis*. JBI.

This online chapter offers practical guidance for conducting a ScR. It is part of a manual whichh is continually updated to reflect any changes to the ScR methodology recommended by the JBI Scientific Committee.

Pham, M. T., Rajić, A., Greig, J. D., Sargeant, J. M., Papadopoulos, A., & McEwen, S. A. (2014). A scoping review of scoping reviews: Advancing the approach and enhancing the consistency. *Research Synthesis Methods*, *5*(4), 371–385.

This journal article assesses the characteristics and range of methodologies used in ScR in different disciplines, including Education and Social Sciences. It highlights levels of rigour and methodological variations in ScR published in English, French, and Spanish.

Pollock, D., Peters, M. D. J., Khalil, H., McInerney, P., Alexander, L., Tricco, A. C., Evans, C., de Moraes, É. B., Godfrey, C. M., Pieper, D., Saran, A., Stern, C., & Munn, Z. (2023). Recommendations for the extraction, analysis, and presentation of results in scoping reviews. *JBI Evidence Synthesis*, *21*(3), 520–532.

This article explains the challenges encountered by researchers conducting ScR. It offers specific guidance on data extraction and analysis as well as on how to present the findings in a ScR.

Tricco, A. C., Lillie, E., Zarin, W., O'Brien, K. K., Colquhoun, H., Levac, D. et al. (2018). PRISMA extension for scoping reviews (PRISMA-ScR): Checklist and explanation. *Annals of Internal Medicine*, *169*(7), 467–473.

This publication introduces a specific PRISMA extension for ScR. It intends to support researchers' reporting of ScR with a more focused guidance on what is methodologically relevant. It includes a checklist and provides examples of good practice.

Westphaln, K. K., Regoeczi, W., Masotya, M., Vazquez-Westphaln, B., Lounsbury, K., McDavid, L., Lee, H., Johnson, J., & Ronis, S. D. (2021). From Arksey and O'Malley and beyond: Customizations to enhance a team-based, mixed approach to scoping review methodology. *MethodsX*, *8*, 101375–101375.

This article presents an enhancement of the Framework by Arksey and O'Malley for ScR with methodological customizations that integrate a team based, mixed methods approach. It recommends the collection of quantitative and qualitative variable for a deeper analysis of the findings.

AI use declaration statement

The authors confirm that no AI tools were used in the preparation, writing, formatting, and editing of this work.

References

Ahmed, A., Al-Khatib, A., Boum, Y. et al. (2023). The future of academic publishing. *Nature Human Behaviour*, *7*, 1021–1026.

Al-Nafjan, E. F., & Alhawsawi, S. (2022). A scoping review of research on tertiary English writing in the Saudi context. *Ampersand*, *9*.

Álvarez, I., Fuertes Gutiérrez, M., & Gallardo Barbarroja, M. (2024). Team teaching in languages: A scoping review of approaches and practices in higher education. *Innovation in Language Learning and Teaching*, *18*(1), 59–77.

Arksey, H., & O'Malley, L. (2005). Scoping studies: Towards a methodological framework. *International Journal of Social Research Methodology*, *8*(1), 19–32.

Aromataris, E., Lockwood, C., Porritt, K., Pilla, B., & Jordan, Z. (Eds.). (2024). *JBI manual for evidence synthesis*. JBI.

Bennett, H., & Brunner, R. (2022). Political and ethical dilemmas in multi-agency participatory research: The role of the buffer zone. *Methodological Innovations*, *15*(3), 387–399.

Chong, S. W. (2025). Synthesis Methods and Reporting Tool (SMART) for research syntheses in Applied Linguistics. *Research Synthesis in Applied Linguistics*, *1*(1), 17–38.

Chong, S. W., & Reinders, H. (2022). Autonomy of English language learners: A scoping review of research and practice. *Language Teaching Research*, *29*(2), 607–632.

Dong, D., & Dong, M. Y. (2024). Exploring the evolution of Applied Linguistics: A bibliometric survey of major research paradigms. In H. Meihami & R. Esfandiari (Eds.), *A scientometrics research perspective in Applied Linguistics* (pp. 11–43). Springer Nature.

Ewald, H., Klerings, I., Wagner, G., Heise, T. L., Stratil, J. M., Lhachimi, S. K., Hemkens, L. G., Gartlehner, G., Armijo-Olivo, S., & Nussbaumer-Streit, B. (2022). Searching two or more databases decreased the risk of missing relevant studies: A metaresearch study. *Journal of Clinical Epidemiology*, *149*, 154–164.

Feng, J., Cho, S., & Luk, G. (2024). Assessing Theory of Mind in bilinguals: A scoping review on tasks and study designs. *Bilingualism: Language and Cognition*, *27*(4), 531–545.

Friend, M., Cook, L., Hurley-Chamberlain, D., & Shamberger, C. (2010). Co-teaching: An illustration of the complexity of collaboration in special education. *Journal of Educational and Psychological Consultation*, *20*(1), 9–27.

Gironzetti, E. & Fuertes Gutiérrez, M. (forthcoming). Publishing. In Lacorte, M. (Ed.), *The Routledge handbook of Hispanic Applied Linguistics* (2nd ed.). Routledge.

Hillman, S., Selvi, A. F., & Yazan, B. (2021). A scoping review of world Englishes in the Middle East and North Africa. *World Englishes*, *40*, 159–175.

Hung, H.-T., Yang, J. C., Hwang, G.-J., Chu, H.-C., & Wang, C.-C. (2018). A scoping review of research on digital gamse-based language learning. *Computers & Education*, *126*, 89–104.

IDESR [International Database of Education Systematic Reviews]. (n.d). https://idesr.org/

Jiang, M. Y. chao, Jong, M. S. yung, Lau, W. W. fat, Chai, C. sing, Liu, K. S. xia, & Park, M. (2020). A scoping review on flipped classroom approach in language education: Challenges, implications and an interaction model. *Computer Assisted Language Learning*, *35*(5–6), 1218–1249.

Law, L. (2024). Application of generative artificial intelligence (GenAI) in language teaching and learning: A scoping literature review. *Computers and Education Open*, 6, 100174.

Li, C., Fryer, L. K., & Chu, S. K. W. (2024). Utilising gamified formative assessment to support English language learning in schools: A scoping review. *Innovation in Language Learning and Teaching*.

Müller, L.-M., Howard, K., Wilson, E., Gibson, J., & Katsos, N. (2020). Bilingualism in the family and child well-being: A scoping review. *International Journal of Bilingualism*, *24*(5–6), 1049–1070.

Peters, M. D. J., Godfrey, C., McInerney, P., Munn, Z., Tricco, A. C., & Khalil, H. (2020). Scoping Reviews. In E. Aromataris, C. Lockwood, K. Porritt, B. Pilla, & Z. Jordan (Eds.), *JBI Manual for Evidence Synthesis*. JBI.

Plonsky, L. (2013). Study quality in SLA: An assessment of designs, analyses, and reporting practices in quantitative L2 research. *Studies in Second Language Acquisition*, *35*(4), 655–687. https://www.jstor.org/stable/26328389.

Plonsky, L. (n.d.). *Bibliography of research synthesis and meta-analysis in Applied Linguistics*. https://lukeplonsky.wordpress.com/bibliographies/meta-analysis/

Rahomati, T., & Sadeghi, K. (2024). Technological innovations used in researching language assessment. A scoping review. In K. Sadeghi (Ed.), *Routledge handbook of technological advances in reseraching language learning* (pp. 118–129). Routledge.

Ripoll, J. C. & Aguado, G. (2014). La mejora de la comprensión lectora en español: Un meta-análisis. *Revista de Psicodidáctica*, *19*(1), 27–44.

RSAL [Research Synthesis in SIGApplied Linguistics]. (n.d.). *Repository of Secondary Research in Applied Linguistics*. https://researchsynthesis.weebly.com/repository-of-secondary-research-in-applied-linguistics.html

Ryan, J. M., Visanich, V., & Brändle, G. (2024). Accommodating and adapting research during the COVID-19 pandemic. In J. M. Ryan, V. Visanich, & G. Brändle (Eds.), *Transformations in social science research methods during the COVID-19 pandemic* (pp. 1–13). Routledge. https://www.taylorfrancis.com/chapters/oa-edit/10.4324/9781032646824-1/accommodating-adapting-research-covid-19-pandemic-michael-ryan-valerie-visanich-gaspar-br%C3%A4ndle.

Satayev, M., Varis, S., Ashirbayev, Y., Koshanova, Z., & Balta, N. (2024). Content and language integrated learning in Kazakhstan: A scoping review. *Journal of Immersion and Content-Based Language Education*.

Tudela-Isanta, A., & Andújar-Molina, O. (2023). La enseñanza del español a personas refugiadas y desplazadas. Revisión de alcance. *Journal of Spanish Language Teaching*, *10*(2), 172–190.

Visonà, M. W., & Plonsky, L. (2020). Arabic as a heritage language: A scoping review. *International Journal of Bilingualism*, *24*, 559–615.

Whitaker, E. M., & Atkinson, P. (2021). *Reflexivity in social research*. Palgrave Macmillan. https://doi-org.libezproxy.open.ac.uk/10.1007/978-3-030-84095-2_3.

Zolfaghari, Z., Karimian, Z., Zarifsanaiey, N., & Farahmandi, A. Y. (2025). A scoping review of gamified applications in English language teaching: A comparative discussion with medical education. *BMC Medical Education*, *25*, 274.

SECTION 7

Systematic review

Section overview

What is a systematic review

Hamish Chalmers
University of Oxford

Definition

Petticrew and Roberts (2006, p. 9) define systematic reviews as "literature reviews that adhere closely to a set of scientific methods that explicitly aim to limit systematic error (bias), mainly by attempting to identify, appraise and synthesise all relevant studies (of whatever design) in order to answer a particular question (or set of questions)".

Systematic reviews stand apart from other (non-systematic) reviews by their attention to completeness and the systematicity and transparency in their conduct and reporting. Systematic reviews make clear every step of the process that led to their compilation. Thus, well-conducted systematic reviews are authoritative and accountable summaries of all (or as much as possible) of the evidence that addresses the question(s) they seek to answer.

Systematic reviews often adopt a narrative approach to synthesis (Popay et al., 2006). They may also present data in tabular form and can also include statistical synthesis of relevant quantative data alongside narrative descriptions of the body of research.

Methodological stages

Systematic reviews can adopt slightly different methodological approaches, but most will involve the following stages.

1. Set the question and define the boundaries

A systematic review should set a clear question it aims to address. Often this will require setting boundaries around things like outcome measures, populations and research design. Questions can be broad, or they can be narrow. A broad ques-

https://doi.org/10.1075/rmal.17.13cha

tion might ask 'What is the extent and nature of research on teaching vocabulary to EFL learners?'. A narrower question might ask 'What are the relative effects of using apps like Duolingo for teaching vocabulary to adult EFL learners compared to traditional paper and pencil techniques on productive vocabulary knowledge?'.

2. Define eligibility criteria

The focus is further defined by specifying the nature of the primary research that will be eligible for the review. Eligibility criteria may include research design, population, intervention or exposure, and outcome measures. For example, to address the second of the two research questions articulated above, the review team might specificy that eligible research must adopt experimental or quasi-experimental designs, be conducted with partipipants aged 18 or over who are learning English outside of formal/compulsory education, and compare the effects of using apps like Duolingo with traditional English teaching approaches on tests of productive vocabulary.

3. Search for eligible evidence

Once eligibility criteria have been defined, a search strategy is developed. This usually means creating a Boolean search string that combines words that one would expect to find in the abstracts of eligible reports, then applying that to appropriate databases. The search strategy can also include citation chaining. This is done by scrutinising the reference sections of eligible studies for other potentially eligible prior research, and by using the 'cited by' functions on some journal websites or Google Scholar, to find potentially eligible later research.

4. Screen for eligibility

Once the initial search has been conducted, the titles and abstracts of all returned reports are screened against the eligibility criteria. Those that cannot be excluded on the basis of the information therein are retained for full text screening, and those that violate any one of the eligibility criteria are discarded. The full reports for all retained abstracts are then sought and their eligibility confirmed (or otherwise) by reading the full texts.

5. Extract data

The data that will form the basis of the synthesis are then extracted from the eligible reports. This will include whatever data the review question demands so that it can be addressed. In experimental research this might include descriptions of

the interventions being compared, the sample, and the outcome measures, as well as the statistical data on the outcomes (such as the n, mean, and standard deviations for each group). In non-experimental research it may include descriptions of the questionnaires, interview schedules, and summaries of the quantitative and qualitative data collected.

6. Appraise quality

A systematic review should not just tell us what a body of evidence says in answer to a research question, but how much confidence we can have in that answer. This is achieved by assessing each report against a standardised quality appraisal tool. Many tools exist to guide this process, and should be chosen based on their appropriateness for the designs of the eligible primary research. Gorard's 'sieve for judging the trustworthiness of research' (Gorard, 2024) is appropriate for experimental and quasi-experimental designs. The Mixed Methods Appraisal Tool (Hong et al., 2018) can be used to appraise the quality of primary research that adopts mixed methods, or for reviews that assemble primary research representing a variety of designs. AXIS (Downes et al., 2016) is a tool designed to guide appraisal of studies that adopt cross-sectional designs. The Joanna Briggs Institute (https://jbi.global/critical-appraisal-tools) provides tools for a variety of different designs.

7. Synthesise

Once all of the above have been conducted, it is time to synthesise the evidence. This means bringing together the eligible evidence to provide an overview and summary in response to the review question. In a narrative synthesis, this involves grouping similar studies together, describing them, reporting on the bottom-line findings and their methodological quality, and arriving at an overall summary of what the evidence tells us. These may be augmented with statistical syntheses, both descriptive and/or analytic.

Software/resources

Rayyan (https://rayyan.ai), Covidence (https://www.covidence.org), and EPPI Reviewer (https://eppi.ioe.ac.uk/eppireviewer-web/home) are all systematic review management systems with varying degrees of complexity. They help consolidate the processes of screening, data extraction, and synthesis. The basic version of Rayyan is free to use, Covidence and EPPI-Reviewer require a subscription.

Examples

Rogde, K., Hagen, Å.M., Melby-Lervåg, M., & Lervåg, A. (2019). The effect of linguistic comprehension instruction on generalized language and reading comprehension skills: A systematic review. *Campbell Systematic Reviews*, *15*(4), e1059.

This is an example of a Campbell systematic review, which have reputations as being exceptionally thorough in preparation, conduct and reporting. This ensures that the method by which conclusions about the body of evidence were arrived at can be fully understood and that the review can be faithfully updated as new evidence comes to light. It combines narrative and statistical syntheses, moderator and sensitivity analyses, risk of bias assessment, and analyses of potential publication bias to paint a comprehensive picture of the state of knowledge on the topic.

Hamilton, C., Schulz, J., Chalmers, H., & Murphy, V.A. (2024). Investigating the substantive linguistic effects of using songs for teaching second or foreign languages to preschool, primary and secondary school learners: A systematic review of intervention research. *System*, *124*, 103350.

This systematic review of the effects of using songs in foreign language teaching uses a multilingual search strategy to increase inclusiveness, assesses the methodological quality of a variety of different designs to provide an holistic view of the overall quality of an eclectic body of research, and provides a narrative synthesis organised by the outcome focuses.

Low, G., & Beverton, S. (2004). A systematic review of the impact of ICT on literacy learning in English of learners between 5 and 16, for whom English is a second or additional language. In *Research Evidence in Education Library*. EPPI Centre, UCL.

While this EPPI-Centre systematic review is now more than 20 years old, it stands as an exemplar of meticulous and systematic methods, and foregrounds user involvement in the production of the review to ensure that it addressed questions that were meaningful to the practitioners whose practice it aimed to inform. It first maps the general field of ICT use in literacy instruction for English learners, then provides in-depth reviews of subcategories drawn from that map.

AI use declaration statement

The author confirms that no AI tools were used in the preparation, writing, formatting, and editing of this work.

References

Downes, M.J., Brennan, M.L., Williams, H.C., & Dean, R.S. (2016). Development of a critical appraisal tool to assess the quality of cross-sectional studies (AXIS). *BMJ Open*, *6*, e011458.

Gorard, S. (2024). Judging the relative trustworthiness of research results: How to do it and why it matters. *Review of Education*, *12*(1), e3448.

Hong, Q.N., Fàbregues, S., Bartlett, G., Boardman, F., Cargo, M., Dagenais, P., Gagnon, M.-P., Griffiths, F., Nicolau, B., O'Cathain, A., Rousseau, M.-C., Vedel, I., & Pluye, P. (2018). The *Mixed Methods Appraisal Tool (MMAT), version 2018* for information professionals and researchers. *Education for Information*, *34*(4), 285–291.

Petticrew, M., & Roberts, H. (2006). *Systematic reviews in the social sciences: A practical guide*. Blackwell.

Popay, J., Roberts, H., Sowden, A., Petticrew, M., Arai, L., Rodgers, M., Britten, N., Roen, K., & Duffy, S. (2006). *Guidance on the conduct of narrative synthesis in systematic reviews: A product from the ESRC Methods Programme*. Lancaster University.

CHAPTER 13

Conducting systematic reviews in Applied Linguistics

A case study and critical reflection

Emily Oxley
University of Glasgow

As the number of systematic reviews in the field of Applied Linguistics increases, so do the expectations around the quality and rigour. However, conducting a rigorous systematic review that is of a high methodological quality involves collaboration, nuanced understanding, and methodological challenges. In this book chapter, I will reflect on the experiences of conducting a systematic review on the effectiveness of language and literacy interventions for children with English as an Additional Language (EAL; Oxley & De Cat, 2019). I discuss some of the main steps of the review process: developing the search strategy, screening and selecting studies, and assessing their quality. I will also share the lessons learned, the difficulties encountered, and reflect on what authors may do differently in the future. These reflections can provide useful insights and guidance for other researchers who are planning or conducting systematic reviews in Applied Linguistics.

Keywords: EAL, language and literacy, intervention, systematic review, methodological reflection

Background

The landscape of research synthesis is undergoing a transformative shift. Whilst once a methodology used largely by the medical field (Hong & Pluye, 2018), it is now expanding across disciplines, with rigorous methodological techniques proving effective in synthesising evidence from a range of sources. Furthermore, advanced software has improved both the speed and rigor of systematic reviews (Bond et al., 2024). Here, I reflect on the process of conducting an older systematic review in language education, first published in 2019 but constituting research undertaken between 2017 and 2018. I consider the difference in methods utilised

https://doi.org/10.1075/rmal.17.13aoxy
Available under the CC BY-NC-ND 4.0 license.© 2026 John Benjamins Publishing Company

in 2017 compared to the ways in which I would tackle such a synthesis now and consider the benefits and challenges of both approaches.

The review in question, *A systematic review of language and literacy interventions in children with English as an Additional Language*, conducted with Professor Cécile De Cat, was first published in 2019 in the Language Learning Journal. It updated Murphy and Unthiah (2015), largely replicating the methodology and synthesised evidence on the effectiveness of language and literacy interventions for children schooled in English, but who spoke a different language at home or in the local community.

The conceptualisation of the review was based on a growing number of children in UK schools speaking *English as an additional language* (EAL). In England, EAL children are classified as those exposed to a language other than English at home (Department for Education, 2018). Globally, terminology varies for children with exposure to multiple languages within an English-speaking education system; terms like English Language Learners (ELL), and English as a Second Language (ESL) are used interchangeably. From a linguistic perspective, these terms are highly problematic, giving little to no indication of the language history, dominance or usage patterns. The UK classification of English being the 'additional' language would suggest that the child's first language (L1) is the language used most frequently outside of the school setting, which is often not the case. Indeed, a child who is 'new to English' (i.e. new to the English-speaking society and school system) will be classified as EAL, as will a child who was born in the English-speaking society, speaks English with siblings and school friends, yet it exposed to another language in the home.

Despite this heterogeneity, the term EAL remains highly used in research and practice. In the English context, teacher generated data (based on teachers' perspectives) were once available about EAL children's English language proficiency, however, the measures were discontinued and now no such data exists (see Strand & Hessel, 2018). While teacher generated data is problematic in itself, these measures were more predictive of children's academic success than the EAL label alone (Strand & Hessel, 2018). Recognising this diversity across children's proficiency is crucial to avoid oversimplifying EAL children's abilities (Hessel & Strand, 2023).

Recent research has sought to address the limitation of a binary descriptive label such as EAL by developing tools that measure multilingual language use (De Cat et al., 2025; Kašćelan et al., 2022), reflecting language knowledge and proficiency as a continuum. However, the binary label EAL persists despite being problematic, as it is already widely used in the research literature and recognised by school staff, who may be the ones with the most need for the research findings.

It is also important to acknowledge that the focus on EAL children's academic proficiency is a highly debated research area. Research into linguistic *deficits* of children from diverse backgrounds is problematic and unethical, emphasising what children cannot do, rather than the strengths that come with multilingualism. However, global trends show lower academic outcomes for EAL children in Maths, English and Science (Murphy, 2021). Regional variation does suggest some schools are better equipped than others to support EAL children; in London, for example, EAL children outperform monolingual peers across multiple subjects, (Strand et al., 2015). Murphy (2021) argues that failing to investigate the causes of poorer outcomes for EAL children in some regions is arguably just as unethical as researching linguistic deficits (Murphy, 2021). Consequently, there is a moral need to understand more about suitable, effective and targeted interventions to bridge the gap between research and practice (Murphy, 2021). Research on EAL learners highlights the importance of early, evidence-based interventions in language and literacy (Biemiller & Boote, 2006). Children who receive targeted early language can sustain progress in language and literacy skills (Bleses et al., 2021) with noted effects for EAL children (Cho et al., 2021). As a result, the current systematic review was conceptualised to find collective evidence on effective interventions to improve the language and literacy skills of school aged children and adolescents with EAL which would be suitable for UK classrooms.

Our review was a conceptual replication of Murphy and Unthiah (2015), therefore much of the original methodology remained. We chose the systematic review methodology of research synthesis over other methodologies to ensure as much similarity to the original review was retained. Since there was already a publication on this topic, a scoping review was deemed unnecessary (for more information on scoping reviews, see Chapters 11 and 12). While a meta-analysis on this topic would be helpful to understand the effectiveness of different interventions on a larger scale (see Chapters 7 and 8 for more information), it was not possible within the time frame and resources available.

Replicating the search terms of Murphy and Unthiah (2015) proved beneficial as a time saving initiative, as there was no need to collaborate with co-workers and agree on a framework (e.g. PICO, SPIDER) and brainstorm potentially useful terms. Using the pre-defined terms meant multiple iterations had already been piloted and the terms accordingly adapted, therefore we had confidence meaningful and relevant results would be returned.

Similarly, search terms had been created with background knowledge of the original first author, who is a well published scholar in this field (e.g. Murphy 2017, Murphy 2018). Consequently, the lead author had knowledge of specific terms which could be subsumed under the language and literacy umbrella, such as 'phonetic decoding' and 'phonetic awareness'.

Example search string

> ("intervention" OR "treatment" OR "program*" OR "implementation")
> AND
> ("minority language" OR "heritage language" OR "additional language" OR "community language" OR "language minority" OR "English language learner" OR "ESL" OR "diverse language" OR "EAL" OR "English as a second language" OR "English as an additional language")
> AND
> ("literacy development" OR "literacy acquisition" OR "literacy skill*" OR "literacy" OR "reading development" OR "reading skill*" OR "reading achievement" OR "literacy achievement" OR "receptive language" OR "receptive vocabulary" OR "productive language" OR "productive vocabulary" OR "expressive language" OR "expressive vocabulary" OR "writing" OR "phonetic decoding" OR "phonetic awareness" OR "bilingualism")
> NOT
> ("case studies" OR "disorder" OR "autism" OR "ethnography")

The collective knowledge of Murphy and Unthiah also framed the inclusion and exclusion criteria. One deviation was the exclusion of studies conducted before 2014, as it was assumed these would have been included in the original 2015 review.

The inclusion criteria were adapted as follows:

- Must include children in education
- Must include a control or comparator group
- Must involve or report on a language or literacy intervention
- Must collect and report on empirical data
- Must analyse progress of EAL learners
- Language of instruction must be English
- Language of the wider community must be English

The exclusion criteria were:

- The studies are not research studies
- There is no language or literacy related intervention
- The learners are of the wrong age
- There is no body of empirical data or systematic analysis
- The learners do not have English as an additional language
- The study is published before 2014
- The learners have a special educational need
- There is no comparator and/or control group
- The learners are in bilingual immersion classrooms.

(Oxley & De Cat, 2019 p. 267)

A further deviation was a reduction in databases searched. Murphy and Unthiah, (2015) searched '*a number of databases*' including Australian Education Index, British Education Index, Linguistics and Language Behavior Abstracts, Education Resource Information Center and Web of Science. In addition, hand searching in key journals in language education was undertaken. Only four of the five stated databases were available to our research team, with no availability to search Australian Education Index. Furthermore, as a result of limited team capacity and time restrictions, there was not scope to carry out hand searches of key journals, which is a clear limitation of this review.

After database searches were carried out, results were exported into an Endnote library, with 2,217 records identified. Following the standard screening procedure used in Cochrane systematic reviews, 25 interventions across 26 studies were included. After qualitative synthesis of included studies, findings pointed to explicit vocabulary instruction and targeted oral language practice, (structured speaking and listening activities) leading to language gains in English, especially for those with lower initial proficiency. Shared reading interventions were also found to be effective when combined with techniques such as the pre-teaching of vocabulary, embedding definitions within texts, or incorporating post-reading reinforcement activities. The review highlighted a significant gap in research, noting a shortage of intervention studies in the UK, a finding replicated from the original review (Murphy & Unthiah, 2015), particularly for adolescents aged 14–18 in upper secondary education. The review emphasised the need for more UK-based research to develop and evaluate effective interventions for children and adolescents with EAL.

Methodological and ethical issues and dilemmas

Anticipated methodological issues prior to beginning the review

The current systematic review was conducted as part of a larger research project about recommendations for primary classrooms in the UK, therefore collective evidence of effective intervention strategies for this specific population in the UK context was needed. Evidence from previous reviews (Oxley, 2019; Murphy & Unthiah, 2015) came mainly from the United States, which, while useful, caused difficulties when providing future recommendations in the UK. In the United States, many English Language Learners are placed in 'immersion' classrooms where the societal language (English) is taught alongside the L1, most often being Spanish (see, for example Hamman-Ortiz, 2020). Resultingly, language specific interventions can be carried out with bilingual teaching staff (e.g. Carlo

et al., 2004; Tong et al., 2008). While this is exemplary practice in many parts of the world, it is infeasible in the UK as there are no bilingual immersion programmes as standard, and oftentimes there is not one dominant L1 within a school (although there are exceptions in some community groups). As the current review was also conceptualised with the purpose of giving recommendations for classroom practice across the UK, papers using bilingual immersion programmes were excluded.

Methodological and ethical issues that arose during the process of synthesis

Defining 'English as an Additional Language'

While the explicit definition for language and literacy interventions was discussed prior to the commencement of searches, at no point did authors explicitly discuss our own definitions of EAL before screening took place. Initially, it was helpful to use England's Department for Education's (DfE) definition mentioned previously, but later at abstract and full text screening, it was clear that more clarity was required.

This is often the case with research synthesis, as much as you can anticipate difficulties, there is a tendency for questions to arise during the screening process. Consequently, it is essential to carry out screening iteratively with colleagues, recording any notes and questions to discuss together before solutions are identified. Even then, any questions raised and outcomes decided upon should be recorded and discussed in the methods section of the review. Indeed, if the review is pre-registered or a protocol is written, these should be updated as decisions are made.

While the DfE definition for EAL seemed to make sense as an inclusion criterion, questions were raised about the extent to which English had to be the main societal language as well as the language of schooling. Cases arose where children were schooled in one language but the main societal language(s) outside of school was different. Our inclusion criteria did specify '*Language of the wider community must be English*' but clarity was needed around the inclusion or exclusion of countries where an 'official' language was used in government, school, or business but societally another language or number of languages were spoken or preferred. In this sense, the definition of EAL set out by the English government is accurate, but the context is entirely different.

Children schooled in an 'official' language but who have preference for the more common societal language will certainly face different challenges to chil-

dren where both the school and societal language differ to that spoken at home. For instance, they will not be compared to monolingual peers in examinations and face potential disadvantage because they have less exposure to the language of their assessment than their classmates. This caused difficulty with one paper in particular, describing an intervention in Lesotho (Van Staden, 2016). The difficulty in this instance was deciding whether children in Lesotho did speak *EAL* in the sense meant by educators in Anglo-centric Western nations, or whether to exclude a potentially effective intervention due to a grey area in terminology. Furthermore, the potential omission of a paper from a non-Western nation raises the wider ethical issue in academic research of Western perspectives dominating the research literature.

Tools and resources available

Since there was little funding allocation for this review, and it was constrained by the resources available at the time, there were practical difficulties which arose during the searching and screening processes. As this research was carried out at a Higher Education Institution, we did have institutional access to research databases for key word searching, which is a barrier that may be encountered by some. However, as previously mentioned, we were constrained by the databases available, meaning we could not fully replicate the original searches of Murphy and Unthiah (2015). In addition, when it came to downloading and screening results, we were directed to use an Endnote library as this was the only available institutionally funded and approved software. Endnote was effective for downloading search results, combining results from different databases and carrying out the process of deduplification. However, it was only then useful for the storage of full texts. There was little scope to effectively use Endnote to screen titles and abstracts and I was instead advised to download all searches into Microsoft Excel and use a spreadsheet to categorise titles and abstracts according to my inclusion and exclusion criteria. This was a somewhat effective technique, however it did come with challenges. Firstly, all screening decisions had to be collated by hand. Excel was programmed to highlight any disagreements between screeners, yet it was an effortful and slow process to find disagreements by hand and calculate inter-rater agreement. This was, however, perhaps a reflection more on the time period when the review was carried out and the use of available tools. With the increase in better technologies for screening and organising search results, there is now less need for the use of Excel in systematic reviewing.

Suitability of tools used for this review

This systematic review was carried out under training from an academic in Health Psychology, utilising the Cochrane approach more typically used for the medical science field (see Chandler et al., 2016 for more information). Consequently, the assessment of paper quality, or critical appraisal (Petticrew & Roberts, 2005), was carried out using a method more appropriate for systematic reviews in the medical field. It is acknowledged that methods used for critical appraisal vary considerably across reviews (Oancea & Furlong, 2007) as well as the purpose of the appraisal. Some reviews use critical appraisal as a judgement criterion for paper exclusion, while others use it to weigh the strength of evidence during synthesis. This review opted to do the latter, as we were aware that there would already be a small pool of studies included, and we did not want to reduce that pool further. However, this again limits the conclusions drawn from the review as some papers were not as rigorous as others.

Many published assessment tools for critical appraisal are available, with elements denoting the appropriateness and effectiveness of the research design. Often, for studies assessing quasi-experimental or randomised control trials, there will be an analysis of the effect of bias on the outcome measures (see Waddington et al., 2017, for further discussion). As such, I decided to use the 'Risk of Bias' (RoB) tool (Chandler et al., 2016) to appraise the quality of included studies. This was another deviation from the original review I was replicating, which had used the Weight of Evidence appraisal (Gough, 2007). However, I was advised the Risk of Bias was an effective and rigorous method to appraise studies in a Cochrane review. This method of bias assessment considered effective selection and randomisation methods, as well as the blinding of participants to the intervention and selective outcome reporting, using a three-point scale 'low', 'medium' and 'high' risk (Chandler et al., 2016).

Two researchers assessed the Risk of Bias for all interventions in this report, however initially no decision was made as to the reconciliation of agreements. At first, we thought reconciliation should be through discussion, but this raised a question of rigour. As the seniority of members of the team differed, discussion alone could have led to the perceived power of the higher status author affecting decisions made. Another problem that arose by using the Risk of Bias tool was the intended field for which the tool was created. Since the RoB is designed for Cochrane reviews, which are typically in the medical field, some of the criteria, such as effective blinding of participants and researchers to condition, were inappropriate to appraise language intervention studies. Some adaptations to the tool were therefore necessary.

Course of action

Definition of key terms

As mentioned previously, an ethical difficulty encountered was the definition of *EAL* as an inclusion criterion. Van Staden's (2016) study was included, even though it can be argued that Lesotho does not fully fit within the inclusion criteria where English must be spoken as part of the 'wider community'. In Lesotho, English is an official language used for government, administration and international communication (Kamwangamalu, 2013). Yet, the main language spoken by the populous is Sotho (or SeSotho) essentially in a monolingual capacity (Lekhotho, 2013). Where the lines blur between EAL or not EAL is when it comes to schooling. The (Se)Sotho language is the main medium for education in Lesotho for the first three years of primary school, however in the fourth year, the language of instruction switches to English, despite it not being spoken by 75% of the population (Lewis, Simons, & Fennig, 2015). Essentially by changing the language of instruction, an almost entire class of English language learners is created, although some children do speak English at home. Based on this knowledge that some children in the class may be able to speak English compared to others who are only exposed during their fourth year of schooling, we decided to include the study in the review. In addition to the intervention possibly providing helpful knowledge about classes where the majority do not speak the medium of instruction, ethically we deemed it important to include studies from countries outside of the typical Western world where research mass occurs.

Tools and resources available

No specific course of action was taken for the lack of screening technology available at the time, as the research team were only able to use the tools readily available through their institution (Excel, Endnote). However, we ensured good note keeping skills were used. Of course, this method is far from optimal and is susceptible to human error. Better solutions exist and will be discussed in the forthcoming section. We acknowledge that many reviews carried out, especially those from students and early career researchers, or those from institutions with less funding, may be restricted in the tools that they use due to a lack of funding for tools which are not open access.

Suitability/appropriateness of tools used

The solution for the mismatch between the specific review and the most appropriate tools was to adapt the Risk of Bias critical appraisal tool (Chandler et al., 2016), only using the assessment criteria which were suitable for studies in education, and calculating the risk based upon this. Furthermore, to add rigour to our review, we added a criterion looking specifically at outcome measures. We found that often, pre- and post-tests were researcher designed, and may not therefore have been validated and norm referenced, which can add additional biases into any claims of participant language or literacy growth. Studies which used only researcher-developed outcome measures with no validity measurements were given a high risk of bias classification for that criterion, whereas the sole use of standardised outcome measures gave studies a low risk of bias. Studies which included a combination of both were thus medium risk of bias. Each author rated the studies independently according to the pre-defined three-point scale (low, medium, high) and inter-rater agreement was calculated from that point. However, inter-rater agreement was not optimal, being 84.6%. Consequently, in order to reconcile these disagreements without power and status issues, we opted for an approach where ratings would be combined, for example, where one reviewer gave a rating of 'high' and a second gave a rating of 'low', we assessed the bias as 'medium'. However, a numerical solution with more than three decision points would have been better, with the use of dispersive statistics to quantify the quality assessment.

Alternative courses of action

While little could be done about the lack of institutional access to Australian Education Index database, an alternative approach to increasing the number of searches returned would be to carry out hand searches, snowballing techniques and to open out the inclusion criteria to include other valuable sources of information such as conference proceedings, reports and the wider grey literature. One clear benefit of using literature outside of commercial publishers is the inclusion of studies which may have null of negative results (Paez, 2017). Reviews which exclude grey literature are more likely to overrepresent significant findings (McAuley et al. 2000; Conn et al. 2003), therefore our review may be artificially inflating the research literature around effective interventions. It is recommended that systematic reviews include unpublished research (Higgins et al., 2019). However, the inclusion of grey literature is not without its challenges and researchers are advised to consider their timeframe, the feasibility of including grey literature and from which sources they will gather these studies. Often, the internet is the main vessel with which researchers locate grey literature, yet it can be difficult

to execute precise searches and search terms tend to be redundant. In addition, the credibility of information sources on the internet can be difficult to verify (Benzies et al., 2006).

We accept that our review may have benefitted from the inclusion of grey literature to ensure results were not subject to publication bias and were truly representative of the effectiveness of language and literacy interventions for children with English as an additional language.

Definition of terms

With regards to the inclusion of studies where English is an official language, but perhaps not a societal language used within the wider community, it would have been beneficial prior to the commencement of screening to decide if an official language was enough to warrant inclusion. As the purpose of this review was to provide recommendations for classroom practitioners in the UK, an alternative route would have been to specifically stipulate that dominant L1 of pupils within the school system nationally needed to be English. Consequently the EAL pupils in question would be directly compared in state examinations to peers who speak English only, therefore a need to improve English language and literacy would be paramount. However, as mentioned previously, this does raise ethical issues around research in education being dominated by that of the Western world, specifically Anglo-centric research.

Alternatively, the review could have specifically targeted pupils who were new to the majority language in the classroom, regardless of what language that is. This could give useful insights from research across the globe and highlight the work of academics outside of the Anglo-centric world. However, this also brings about problems with generalisability. Again, with the overall aims of the paper to recommend effective language and literacy interventions for practitioners in UK schools, it may be redundant to promote collective evidence from interventions in other languages, which may have fundamental differences, such as different alphabetical systems and language transparency among others.

Tools and resources available

While opting to use the Cochrane verified critical appraisal tool due to advice around rigour (Risk of Bias; Chandler et al., 2016), in hindsight, it may have been a better option to choose tools in the first instance which were relevant for studies conducted in language education, rather than spending time making adaptations and trialling these over multiple iterations. There is a plethora of appraisal tools available for use in evidence synthesis, which could have been more appropriate

for this review. I have since trialled the Quality Assessment Tool of Studies with Diverse Designs (QATSDD: Sirriyeh et al., 2012), but again, being written primarily with the health sciences in mind, adaptions were necessary to review studies which took place outside of a clinical setting.

More recently, I have started to use the Mixed-Methods Appraisal Tool (MMAT, Hong et al., 2018) for reviews in language education. This tool is valued for its flexibility with studies of differing natures and this can be helpful in intervention studies which do not adopt a true randomised control design (as they are, for example, taking place at classroom or school level, thus are more frequently classed as quasi-experimental). Our inclusion criteria for this review did specify empirical, quantitative studies, which is why, at the time, we believed the Risk of Bias to perhaps be the most appropriate. It is clear that there is more work that can be done around developing the best quality assessment tools in the field of education broadly, and specifically language education.

Tool availability

By using hand techniques for data organisation there was sometimes difficulty with data management and meticulous notes needed to be kept at all times. Screening and agreement calculations could have been managed more optimally through use of evidence synthesis software, had it been available to us at the time. An example of such a tool is EPPI-Reviewer (Thomas et al., 2022) which can be used to manage searches and adds automisation to the process. This ensures papers are sufficiently screened across the review team, recording the frequency of reviewer decisions and calculating inter-rater agreement through Cohen's Kappa, giving a far clearer indication of agreement than simple percentages. These semi-automisation tools can also go one step further to increase the rigour of reviews. Firstly, they can link with tools to use snowballing techniques to database searches, reducing the chance that papers may be missed due to search terms and/or database availability. Furthermore, with the current technological advances we are seeing, they offer scope to use artificial intelligence to help with the often long and laborious screening processes. For example, a recent addition to EPPI-Reviewer is their priority screening function (Thomas et al., 2022) which uses machine learning during the screening process to predict which papers are more or less likely to be included. This process is based on training data from the initial screening stage, shuffling the papers most suitable to the top of each screener's list, improving overall time efficiency. This can be particularly useful for large scale reviews or meta-reviews where recording processes by hand becomes infeasible.

The use of AI in systematic reviews offers notable efficiencies but comes with several important caveats. For instance, AI may lead to methodological inconsistencies, applying inclusion criteria or quality assessment unevenly across studies. It may also misclassify studies or extract data inaccurately, particularly when dealing with subtle differences across criteria. Furthermore, an overreliance on AI can reduce researchers' critical engagement with the literature, potentially weakening the rigour of the review. AI should therefore be used as a support tool rather than replacing human judgement, with clear documentation of its role, validation of all outputs by reviewers, and careful adherence to transparency standards.

Conclusion

Overall, the process of undertaking this systematic review was an opportunity to learn not only about language and literacy interventions for EAL children, but also to advance my knowledge of this methodology. Firstly, I learned the importance of clarity of key terms and criteria prior to undertaking a review. However, I acknowledge that in the majority of reviews, questions are raised at every stage of the process. Regular meetings with the research team are therefore recommended, with feedback around any arising conceptualisation or methodological issues. These issues should clearly be documented and any deviations made from the research protocol should be acknowledged in any following publications.

In terms of organisational tools for the screening process, I now ensure that all reviews I undertake use appropriate technology to assist in the process. In the time that has passed since this review was published, more tools have become available and I recommend future authors to research the best tools for their specific needs. Inequalities do exist in the availability of screening tools, with some institutions directing students to free resources over those with fees, and this is a reality of the higher education landscape. Without additional research funding, some projects cannot utilise the most appropriate resources due to budgeting constraints. However, there are freely available alternatives, such as Rayyan (Ouzzani et al., 2016) which may not offer the same level of capabilities but certainly add to the rigour of screening and record keeping. Other scholars have opted to carry out specific coding packages (see e.g. Westgate, 2019) although of course a barrier to this is the ability to use the precise coding techniques necessary.

Similarly, while many tools now exist to assess methodological quality and rigour of systematic reviews with different scopes, there still appears to be a lack of specific tools available to appraise papers in language education, and I encourage the development of such in future.

As the educational landscape changes, and the number of evidence synthesis papers increases in Applied Linguistics, so do the tools available for use in our field. The SMART reporting tool for Applied Linguistics (Chong, 2025), for example, offers researchers a helpful framework during the conceptualisation of their review. Furthermore, the improvements in research synthesis practices in our field have led to more rigorously completed systematic reviews. The increase in published protocols for example aids reviewers in the earliest stages of their review, by undergoing a peer review process where key decisions such as the database choice, search terms and inclusion and exclusion criteria can be critiqued by experts in the field and changed based on recommendations. The use of published protocols adds a level of trust to the collective evidence gathered and conclusions drawn, and aids in future replications. Indeed, some journals now in research synthesis insist on published protocols before considering synthesis papers (see Chapter 4).

To conclude, a number of lessons have been learned both during the process of this review and in the subsequent years. Future researchers in Applied Linguistics should consider the points brought forward and ensure they have a clear understanding of both the ethical and methodological issues which may arise when undertaking a review in this field of study.

Questions for discussion

1. How can we endeavour as research synthesis methodologists to ensure search results are not dominated by the Western world?
2. What is the value of the inclusion of grey literature? Does the value added mitigate the challenges it brings?
3. How do we work towards an equitable system so that early career researchers are trained effectively in the best and most appropriate methods and tools for their field?

Further reading



Chalmers, H., Faitaki, F., & Murphy, V. (2023). Setting research priorities for English as an additional language: What do research users want from EAL research? *Language Teaching for Young Learners*, *6*(1), 5–31.

This paper reports on a Priority Setting Partnership (PSP) aimed at identifying research priorities in the field of EAL. Teachers, parents and students collaborated to identify the most pressing research questions. Through an online questionnaire and a workshop, the researchers

consolidated and ranked these questions into a list of ten priorities, focussing on effective teaching strategies, support for EAL learners with special educational needs, and the socioemotional development of EAL students. Additional priorities related to language proficiency and academic achievement, the benefits and challenges of bi- and multilingualism for EAL students and the need for further longitudinal studies.

doi Fromm, Y.M., Martin, F., Gezer, T., & Ifenthaler, D. (2025). Best practices for conducting systematic reviews: Perspectives of experienced systematic review researchers in educational sciences. *Technology, Knowledge and Learning*, *30*, 1–28.

This paper explores evolving methodologies for conducting systematic reviews in education to enhance their quality and effectiveness. Best practices across systematic review stages (e.g. screening/coding) are discussed following interviews with experienced reviewers. Key themes include the importance of a transparent research process, the benefits and challenges of systematic reviews, team collaboration, ethical considerations, and the role of technology.

doi Stewart, M., Skinner, B., Hou, H., & Kelly, R. (2024). A systematic literature review of home-school partnership for learners with English as an Additional Language (EAL): A way forward for the UK and Ireland. *Irish Educational Studies*, *43*(2), 301–327.

This systematic review investigates the collective evidence base for home-school partnerships to enhance the educational experience of EAL children. The study identifies international trends and gaps in the UK and Ireland, highlighting the importance of effective communication, parental involvement, and culturally responsive practices. The review reveals a lack of school-directed interventions in the UK and Ireland and emphasises the need for schools to engage more actively with EAL families. Enhancing home-school partnerships can significantly benefit EAL learners by addressing their unique needs and promoting their academic and socioemotional development.

AI use declaration statement

During the preparation of this work, the author used Co-Pilot and ChatGPT (GPT-4o) to suggest word count reductions while preparing the final manuscript.

As author, I confirm that after using the tools, I reviewed and edited the content as needed and take full responsibility for the content of the publication, in alignment with COPE and the publisher's AI policy.

References

doi Bleses, D., Dale, P. S., Justice, L., Højen, A., Vind, B. D., et al. (2021). Sustained effects of an early childhood language and literacy intervention through second grade: Longitudinal findings of the SPELL trial in Denmark. *PLOS ONE 16*(10): e0258287.

doi Benzies, K.M., Premji, S., Hayden, K.A., & Serrett, K. (2006). State-of-the-evidence reviews: advantages and challenges of including grey literature. *Worldviews on Evidence-Based Nursing*, *3*(2), 55–61.

Biemiller, A., & Boote, C. (2006). An effective method for building meaning vocabulary in primary grades. *Journal of Educational Psychology*, *98*(1), 44–62.

Bond, M., Khosravi, H., Bergdahl, N., Buntins, K., De Laat, M., Oxley, E., Chong, S. W., Negrea, V. & Händel, M. (2024). Digital evidence synthesis tools in educational technology research: A systematic mapping review. *ResearchGate*.

Carlo, M. S., August, D., McLaughlin, B., Snow, C. E., Dressler, C., Lippman, D. N., Lively, T. J., & White, C. E. (2004). Closing the gap: Addressing the vocabulary needs of English-language learners in bilingual and mainstream classrooms. *Reading Research Quarterly*, *39*(2), 188–215.

Chandler, J., Clarke, M., McKenzie, J., Boutron, I., & Welch, V. (2016). Cochrane methods. *Cochrane Database Syst Rev*, *10*(Suppl 1), 29.

Cho, Y., Kim, D., & Jeong, S. (2021). Evidence-based reading interventions for English language learners: A multilevel meta-analysis. *Heliyon*, *7*(9). https://www.cell.com/heliyon/fulltext/S2405-8440(21)02088-0.

Chong, S. W. (2025). Synthesis Methods and Reporting Tool (SMART) for research syntheses in Applied Linguistics. *Research Synthesis in Applied Linguistics*, *1*(1), 17–38.

Conn, V. S., Valentine, J. C., Cooper, H. M., & Rantz, M. J. (2003). Grey literature in meta-analyses. *Nursing Research*, 52, 256–261.

De Cat, C., Gusnanto, A., Kašćelan, D., Prévost, P., Serratrice, L., Tuller, L., & Unsworth, S. (2025). How detailed do measures of bilingual language experience need to be? A cost–benefit analysis using the Q-BEx questionnaire. *Bilingualism: Language and Cognition*, 1–12.

Department for Education. (2018). *Schools, pupils and their characteristics* (pp. 1–15). Government UK. https://assets.publishing.service.gov.uk/media/5b31eb0840f0b67f7f306124/Schools_Pupils_and_their_Characteristics_2018_Main_Text.pdf

Gough, D. (2007). Weight of evidence: A framework for the appraisal of the quality and relevance of evidence. *Research Papers in Education*, *22*(2), 213–228.

Hamman-Ortiz, L. (2020). Becoming bilingual in two-way immersion: Patterns of investment in a second-grade classroom. *International Journal of Bilingual Education and Bilingualism*, *26*(1), 69–83.

Hessel, A. K., & Strand, S. (2023). Proficiency in English is a better predictor of educational achievement than English as an Additional Language (EAL). *Educational Review*, *75*(4), 763–786.

Higgins, J. P. T., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M. J., & Welch, V. A. (Eds.). (2019). *Cochrane handbook for systematic reviews of interventions* (2nd ed.). John Wiley & Sons.

Hong, Q. N., Pluye, P., Fàbregues, S., Bartlett, G., Boardman, F., Cargo, M., Dagenais, P., Gagnon, M.-P., Griffiths, F., Nicolau, B., O'Cathain, A., Rousseau, M.-C., & Vedel, I. (2018). *Mixed Methods Appraisal Tool (MMAT), version 2018*. Canadian Intellectual Property Office.

Hong, Q. N., & Pluye, P. (2018). Systematic reviews: A brief historical overview. *Education for Information*, *34*(4), 261–276.

Kamwangamalu, N. (2013). Language-in-education policy and planning in Africa's monolingual kingdoms of Lesotho and Swaziland. In J. Tollefson (Ed.), *Language policies in education: Critical issues* (2nd ed., pp. 156–171). Routledge.

Kašćelan, D., Prévost, P., Serratrice, L., Tuller, L., Unsworth, S., & De Cat, C. (2022). A review of questionnaires quantifying bilingual experience in children: Do they document the same constructs?. *Bilingualism: Language and Cognition*, *25*(1), 29–41.

Lekhotho, M. (2013). Lesotho: Organization, structures and challenges. In C. Harbor, (Ed.), *Education in Southern Africa* (pp. 55–80). Bloomsbury.

Lewis, P., Simons, G., & Fennig, C. Eds. (2015). *Ethnologue: Languages of the World* (18th ed.). SIL International.

McAuley, L., Pham, B., Tugwell, P., & Moher, D. (2000). Does the inclusion of grey literature influence estimates of intervention effectiveness reported in meta-analyses? *The Lancet*, *356*, 1228–1231.

Murphy, V.A. (2017). Literacy development in children with English as an additional language (EAL). In J. Enever & E. Lindgren (Eds.), *Early language learning: Complexity and mixed methods* (pp. 41–61). Multilingual Matters.

Murphy, V.A. (2018). Literacy development in linguistically diverse pupils. In D. Miller, F. Bayram, J. Rothman, & L. Serratrice (Eds.), *Bilingual cognition and language: The state of the science across its subfields* (pp. 155–182). John Benjamins.

Murphy, V.A. (2021). 5 Social justice and questions of marginalization in research with linguistically diverse children. In A. Pinter & K. Kuchah (Eds.), *Ethical and methodological issues in researching young language learners in school contexts* (pp. 87–105). Multilingual Matters.

Murphy, V.A., & Unthiah, A. (2015). *A systematic review of intervention research examining English language and literacy development in children with English as an Additional Language (EAL)*. Educational Endowment Foundation. http://www.naldic.org.uk/Resources/NALDIC/Research%20and%20Information/Documents/eal-systematic-review-prof-v-murphy.pdf

NALDIC. (2014). *The national audit of English as an additional language training and development provision*. National Association for Language Development in the Curriculum. https://www.naldic.org.uk/Resources/NALDIC/Research%20and%20Information/Documents/NALDIC%20-%20EAL%20Audit%202014%20FINAL%20FINAL%20YF%20Oct%2014.pdf

Oancea, A., & Furlong, J. (2007). Expressions of excellence and the assessment of applied and practice-based research. *Research Papers in Education*, *22*(2), 119–137.

Ouzzani, M., Hammady, H., Fedorowicz, Z., & Elmagarmid, A. (2016). Rayyan — A web and mobile app for systematic reviews. *Systematic Reviews*, *5*(210).

Oxley, E. (2019). Word learning and vocabulary development in children with English as an additional language. [Unpublished doctoral thesis]. University of Leeds.

Oxley, E., & De Cat, C. (2019). A systematic review of language and literacy interventions in children and adolescents with English as an additional language (EAL). *The Language Learning Journal*, *49*(3), 265–287.

Paez, A. (2017). Gray literature: An important resource in systematic reviews. *Journal of Evidence-Based Medicine*, *10*(3), 233–240.

Petticrew, M., & Roberts, H. (2005). *Systematic reviews in the social sciences: A practical guide*. Wiley & Sons.

Sirriyeh, R., Lawton, R., Gardner, P., & Armitage, G. (2012). Reviewing studies with diverse designs: The development and evaluation of a new tool. *Journal of Evaluation in Clinical Practice*, *18*, 746–752.

Strand, S., & Hessel, A. (2018). *English as an additional language, proficiency in English and pupils' educational achievement: An analysis of local authority data*. The Bell Foundation. https://www.bell-foundation.org.uk/wp-content/uploads/2018/10/EAL-PIE-and-Educational-Achievement-Report-2018-FV.pdf

Strand, S., Malmberg, D. L., & Hall, D. J. (2015). *English as an Additional Language (EAL) and educational achievement in England: An analysis of the National Pupil Database*. Department of Education.

Thomas, J., Graziosi, S., Brunton, J., Ghouze, Z., O'Driscoll, P., Bond, M., & Koryakina, A. (2022). *EPPI-Reviewer: Advanced software for systematic reviews, maps and evidence synthesis*. EPPI-Centre, UCL Social Research Institute, University College London. https://eppi.ioe.ac.uk/cms/Default.aspx?tabid=2914

Tong, F., Lara-Alecio, R., Irby, B., Mathes, P., & Kwok, O. M. (2008). Accelerating early academic oral English development in transitional bilingual and structured English immersion programs. *American Educational Research Journal*, *45*(4), 1011–1044.

Van Staden, A. (2016). Reading in a second language: Considering the "simple view of reading" as a foundation to support ESL readers in Lesotho, Southern Africa. *Per Linguam*, *32*(1), 21–40.

Waddington, H., Aloe, A. M., Becker, B. J., Djimeu, E. W., Hombrados, J. G., Tugwell, P., Wells, G., & Reeves, B. (2017). Quasi-experimental study designs series — Paper 6: Risk of bias assessment. *Journal of Clinical Epidemiology*, *89*, 43–52.

Westgate, M. J. (2019). Revtools: An R package to support article screening for evidence synthesis. *Research Synthesis Methods*, *10*(4), 606–614.

CHAPTER 14

Challenges and practical solutions in Applied Linguistics systematic review

A reflective perspective

Takumi Aoyama
Meiji University

With the expansion of Applied Linguistics as a research field and the broadening of its topics, the synthesis of accumulated research has become increasingly active over the past decade. This trend is facilitated by the growing number of academic journals, especially open-access journals, being published worldwide. As a result, various guidelines for conducting systematic reviews (e.g., Chong, 2025; Page et al., 2021) have been established in Applied Linguistics, as in other disciplines. These guidelines have advanced the discussion on methods for data collection, the establishment of inclusion/exclusion criteria, and the selection of analytical approaches. It is generally understood that most researchers conduct their reviews in accordance with these established guidelines. This chapter, however, explores the practical challenges encountered while conducting a systematic review of L2 motivation research in the Japanese EFL context. Based on my own experience, I describe dilemmas related to language choice, database selection, and the thoroughness of the review. The chapter aims to encourage broader discussions on how to effectively conduct systematic reviews in a field with increasingly diverse and multilingual nature of Applied Linguistics research, and provides a reflective perspective on these issues and proposes solutions for future systematic reviews.

Keywords: systematic review, research synthesis, language bias, country-in-focus, L2 motivation

1. Background

Systematic reviews have become a major trend in the field of Applied Linguistics, with a proliferation of studies that analyze research trends across various domains in the past decade. In this chapter, drawing on my own experience of conducting

https://doi.org/10.1075/rmal.17.14aoy
Available under the CC BY-NC-ND 4.0 license.© 2026 John Benjamins Publishing Company

a systematic review (Aoyama, 2026), I will describe the challenges encountered during the process of a review on L2 motivation research, along with the corresponding strategies and solutions that were implemented. This chapter draws on a systematic review of language learning motivation research in the Japanese EFL educational context. This systematic review served as the literature review for a PhD study that investigates the motivational characteristics of Japanese EFL learners and the complex, dynamic changes in their motivation for learning English. Through this chapter, I aim to offer a perspective on potential practical challenges — such as database selection, exhaustiveness, and language biases — and to propose possible solutions for future systematic reviews in the field.

1.1 Background of review: L2 motivation

In 2017, when this systematic review was conducted, systematic reviews published in L2 motivation research were Masgoret and Gardner's (2003) meta-analysis of studies conducted in the tradition of socioeducational model, and Boo et al.'s (2015) systematic review of studies published on the topic of L2 motivation between 2005 and 2014. After 2020, more systematic reviews on L2 motivation research have been published by researchers based in various research settings and with diverse perspectives. As it is not feasible to mention all of these studies due to space limitations, a selection of recent works is presented in Table 1.

The systematic review introduced in this chapter is a study as a part of literature review for a PhD project (Aoyama, 2026). The key rationale for conducting a systematic review was to address limitations of previous global reviews, such as the systematic review by Boo et al. (2015) and explore a broader landscape of L2 motivation research in the Japanese educational context. In their review, Boo et al. (2015) included studies published only in international, high-profile English-language journals. While their review identified an increase in research in Asian contexts, with 38 studies from Japan (11.34% of their sample), the review by Aoyama's (2026) had a more specific focus on the research conducted in the Japanese context. This expanded search included studies published in both English and Japanese from refereed and non-refereed journals, as well as book chapters. Furthermore, while the Boo et al.'s (2015) review was limited to L2 motivation publications from 2005 to 2014, a period of ongoing growth in the field, Aoyama's (2026) study not only aimed to partially replicate this timeframe but also included subsequent publications to allow for a comparison of research trends.

Table 1. Recent systematic reviews of L2 motivation research

Study	Focus of review
Boo et al. (2015)	The origins and nature of the unprecedented surge in L2 motivation publications between 2005 and 2014. Theoretical and research methodological trends.
Aryadoust et al. (2024)	State of research on LLM, including a thematic review of variables affecting LLM and a reliability generalization meta-analysis of measurement tools.
Güleçoğlu and Öztürk (2021)	Research perspectives on EFL learning motivation in the Turkish context. How EFL motivation has been studied as a research matter.
Shurovi et al. (2025)	Research trends in language learning strategies (LLS) in relation to learner motivation.
Mahmoodi and Yousefi (2022)	Synthesis of scholarship on L2 motivation, examining theoretical approaches, methods, instruments, and learning contexts in recent research.
Al-Hoorie et al. (2025)	Self-determination mini-theories in second language learning.
Wu (2022)	Review of publications on L2 language learning motivation using a bibliometric approach. Examined changes in topics.

1.2 Approach to review

Aiming to review research trends of L2 motivation within the Japanese educational context, a systematic review was conducted, with a particular focus on participant groups, theoretical frameworks, and methodological characteristics. The review followed the approach used by Boo et al.'s (2015) review, with modifications to enable an in-depth search of literature, including papers in Japanese local scholarly databases. First, database search was performed. For initial article search, scholarly databases including CiNii, J-STAGE (Japanese national scholarly databases), ERIC, LLBA, MLA, and PsycINFO were searched with the keywords "((L2 motivation) OR (language learning motivation) OR (English motivation)) AND ((Japan) OR (Japanese))", as well as the Japanese translation of these keywords in article title and abstract (see Section 4.2 for details). As a result, the initial search yielded 600 papers. Also, a manual search of anthologies on L2 motivation research (Table 2) and ancestry search of papers in the references list in those anthologies were conducted, and 48 papers and book chapters were added to the article pool.

Subsequently, the papers retrieved in the initial database search were screened with the following criteria: (1) the study deals with motivation to learn English

Table 2. Anthologies included in the manual search

Editors	Title
Dörnyei & Ushioda (2009)	*Motivation, language identity and the L2 self*
Apple et al. (2013)	*Language learning motivation in Japan*
Csizér & Magid (2014)	*The impact of self-concept on language learning*
Dörnyei et al. (2015)	*Motivational dynamics in language learning*
MacIntyre et al. (2016)	*Positive psychology in SLA*
Apple et al. (2017)	*L2 selves and motivations in Asian context*

in Japan; (2) the study was published between 2005 and early 2017; (3) study was published in a journal (refereed or non-refereed) or as a book chapter; and (4) the study was published either in English or Japanese. As conference presentations and dissertations are likely to be published in books or journals and overlap with papers that meet the inclusion criteria, studies in conference abstracts, proceedings, and dissertations were excluded.

After the screening based on the papers' titles, abstracts, and contents, as well as the removal of overlapped records, 401 papers and book chapters were excluded from further analyses, and 247 cases were included in the final article pool. This contrasts significantly with the 38 Japan-based studies found by Boo et al.'s (2015) systematic review which serves as a reference case to the present review of my own. This discrepancy attributes not only to the pure increase in the number of publications on language learning motivation in the Japanese research context due to the growing interest in the topic of L2 motivation in Japan, but also to the differences in the databases used in each study. While Japan-based studies included in Boo et al.'s (2015) review were the ones that were published in international journals, Aoyama's (2026) review covered publications in Japanese local journals as well.

The 247 papers were then coded based on 10 components. These included publication language, year, type of paper (empirical/theoretical), focus (Motivation, Motivating, Demotivation), demographics (Primary, Lower secondary, Upper secondary, University+, Other, Mixed), theoretical framework (e.g., Gardner, L2MSS, SDT, CDST, Unclear, More than one), method (Quantitative, Qualitative, Mixed-method), data collection tools (Questionnaires, Interviews, Observations), number of data collections (Once, Twice, Three times, More than three), and quantitative analysis techniques (e.g., Clustering, t-test/ANOVA, Correlation, Regression, EFA, SEM, Descriptive only).

1.3 Key findings

1.3.1 *Publication language*

By deliberately expanding the scope beyond the limitations of previous global reviews, such as the influential work by Boo et al. (2015) which predominantly focused on high-profile, English-language international journals, Aoyama's (2026) review revealed a more comprehensive overview of L2 motivation research within the Japanese educational context. Compared to the number of Japan-based language learning motivation studies reported in Boo et al.'s (2015) study ($k = 38$), this study identified 247 papers as a result. Among the 247 papers, 77 (31.17%) were written and published in Japanese, while 170 of them (68.83%) were published in English. This indicates that more than 200 studies were not included in Boo et al.'s (2015) analysis of international journals, and the article pool of Aoyama's study reflects a more detailed picture of trends in language learning motivation research in the Japanese educational context. This indicates that the selection of publication language in the article pool needs to be carefully considered when planning a systematic review.

1.3.2 *Target demographics*

This comprehensive mapping was instrumental in critically identifying significant gaps in the existing literature on L2 motivation in Japan. For instance, the review revealed an overwhelming concentration of research on a particular group of learners. While Boo et al. (2015) found that upper secondary students were the most frequently studied group globally (51.64%) followed by university students (22.69%), Aoyama's (2026) review revealed that research in Japan is heavily concentrated on learners at the university level and above (73.54%). This suggests that, in the Japanese research context, there is a lack of understanding regarding the motivational trajectories of primary (9.42%), lower secondary (4.04%), and upper secondary (8.07%) students — a demographic Boo et al. (2015) found to be most studied globally.

1.3.3 *Focus of papers and theoretical frameworks*

Similar to Boo et al.'s (2015) finding that 67% of studies focused on "Motivation", review in Aoyama (2026) found that the overwhelming majority of papers in the Japanese context (72.87%) focused on "Motivation" itself, with fewer addressing "Motivating" (22.27%) or "Demotivation" (4.45%). Also, both reviews noted the emergence and increasing use of newer theoretical concepts. Self-Determination Theory (SDT) was the most frequent framework used in the Japanese context (24.38%). The L2 Motivational Self System (L2MSS) showed a notable increase, particularly between 2011 and 2012, similar to its emergence noted by Boo et al.

(2015). Complex and Dynamic Systems Theory (CDST) also emerged between 2012 and 2017. A significant percentage (22.73%) of studies used "More than one concept".

1.3.4 *Research method*

Regarding data collection methods used in L2 motivation research, Aoyama (2026) highlighted a strong preference in the Japanese context for using questionnaires (79.63%) and relying on one-shot data collection (64.29%). This suggests that many studies capture a static snapshot of motivation rather than its dynamic nature over time. Also, in terms of research methods, Boo et al. (2015) identified a growth in qualitative and mixed-methods approaches globally. Aoyama (2026) confirmed this trend within Japan, finding that while quantitative methods remained the most common overall (60.65% of empirical studies), there was a shift from 100% quantitative studies in 2005 to a more balanced distribution with qualitative and mixed methods accounting for approximately 50% of empirical studies by 2017. While the emergence of diverse theoretical frameworks such as SDT, L2MSS, and CDST, and an increase in qualitative and mixed-methods studies were noted, mirroring global trends, the methodological preference for cross-sectional designs limited insights into motivational evolution and its interplay with contextual factors. Thus, Aoyama's (2026) review pointed to a critical need for research capable of capturing the complex and dynamic changes in learner motivation over time, rather than static snapshots in the Japanese research context.

2. Methodological and ethical dilemma in the systematic review

This systematic review (Aoyama, 2026) was conducted to address the lack of a comprehensive literature review on L2 motivation research within the Japanese educational context. Throughout the review process, clear similarities and differences emerged between global reviews (e.g., Boo et al., 2015) and those focusing on a more specific research context. The clear discrepancies in the breadth and depth of findings between Aoyama (2026) and Boo et al. (2015), largely attributable to the differing criteria for literature inclusion, naturally lead to the subsequent discussion. The significant contrast underscores the profound impact of decisions regarding the selection of target publications in systematic reviews within Applied Linguistics, particularly concerning the inclusion of English-only, top-tier journal articles versus a broader inclusion of non-English and more locally focused research.

Reflecting on the research process raises a fundamental question: if the trends in findings differ between global and more focused reviews, what does this imply

about the significance and role of systematic reviews? This review, in particular, brought to light concerns about the validity of its findings for a global readership and questions regarding study quality. In this section, those methodological considerations, as highlighted by the differing outcomes, will be critically examined through my own reflections on the research process and dilemmas I encountered during the research process.

2.1 Dilemma 1: Language choice in systematic reviews

The first point was the choice of publication language in the review. Since the review aimed to explore the characteristics of L2 motivation research in Japan, in order to set up research questions for a larger-scale study, as mentioned earlier, the review included papers written in both English and Japanese. Thus, the concern was that to what extent the review should maintain the methodological rigor, while embracing diversities and exploring research trends in a particular context.

The first question is whether limiting publication language in systematic reviews is a valid approach. Language bias, as one of the non-reporting biases, is defined as "the publication of research findings in a particular language, depending on the nature and direction of the results" in the Cochrane guideline (Boutron et al., 2019, p.182). Stern and Kleijnen (2020) argue that limiting systematic reviews to studies published only in English is a common practice that introduces language bias. The most frequent reasons cited by authors for this restriction are a lack of time, funding, and access to professional translators. This exclusion risks ignoring key data, missing important cultural contexts, and ultimately limiting the usefulness of a review's findings. The authors highlight that major research synthesis organizations such as JBI, the Campbell Collaboration, and Cochrane all recommend that searches should not be restricted by language. They conclude that by limiting a review to a single language, researchers violate the core purpose of a systematic review, which is to make informed decisions based on the best available evidence.

The issue of language bias in systematic review is a recognized concern, as restricting study inclusion to a single language is a common practice that may introduce bias and omit key data. However, empirical studies examining the effects of language choice in data collection have yielded mixed results. For example, a systematic review by Morrison et al. (2012) found no overall evidence of systematic bias from language restrictions in systematic reviews of conventional medicine, and several studies included in that review reported no significant differences in summary treatment effects between English-only and language-inclusive meta-analyses.

An analysis by Jüni et al. (2002) also concluded that excluding non-English trials generally has a limited effect on summary treatment estimates. In contrast, other findings suggest a more definitive impact. The same study by Jüni et al. found that trials in languages other than English were more likely to show statistically significant results and reported treatment effects that were, on average, 16% more beneficial than those in English-language trials. Further complicating the matter are conflicting findings on methodological quality; some research indicates English-language trials have higher methodological quality, while other studies have found only minor differences. Ultimately, while including studies in other languages can improve the precision of pooled estimates, the overall effect of language restriction is difficult to predict for individual reviews and can vary significantly across different medical fields.

However, these discussions on language bias occur in fields such as health care and epidemiology, and primarily center on the inclusion of clinical trials in systematic reviews and meta-analyses, and the effects of language bias in systematic review are not explored in Applied Linguistics research. From an ethical standpoint, decisions regarding the level of inclusivity in systematic review are not simply technical procedures, but are closely related to the researcher's reflexivity (e.g., Suri & Clarke, 2009). Therefore, it is crucial to carefully consider how language bias manifests in the context of Applied Linguistics research. Also, we must consider how researchers can present and discuss the rationales behind language choices from the viewpoint of researchers' reflexivity.

2.2 Dilemma 2: Providing details of non-English publications

The second point, which is connected to the second concern, is the extent to which details of studies published in languages other than English — in this case, Japanese — should be included iin the analysis process. While it is common practice to list the studies included in a systematic review's dataset as an appendix, merely providing bibliographic information is often insufficient for readers to access the original articles and grasp their specific details.

This issue becomes particularly critical for research topics that require a deep understanding of a specific context, such as a review of studies on language learners within a particular country. To ensure sufficient clarity, measures must be taken to make the content of the original studies comprehensible to readers who are not speakers of the source language.

For instance, in their review of English Language Teaching (ELT) research in Japan, Terasawa and Segawa (2025) address this problem. They argue that an English-only selection policy for their review would lead to significant selection bias. They summarize that Japanese academia, especially in the humanities and

social sciences, has a long tradition of prioritizing publication in Japanese. This tendency is particularly strong when discussing local or context-specific knowledge, such as educational policies and domestic issues, as the primary audience is typically Japanese-speaking.

2.3 Dilemma 3: Thoroughness of review

The third dilemma I had in the study was the thoroughness of the review. While it would be ideal to include as many studies as possible in the dataset, it would be almost impossible to include all studies in the field from different research settings. According to SJR's journal rankings, there are 1,347 journals listed as Q1–Q4 journals within the topic of "Linguistics and Language". Since the main aim of my own systematic review was to cover recent theoretical and methodological trends of L2 motivation research in Japan, limiting the publication period and the range of journals was the choice in the actual study.

In previous systematic reviews of L2 motivation research, it is common for studies to limit both the publication period and the range of journals. For example, a systematic review by Mahmoodi and Yousefi (2022) focused on publications between 2010 and 2019. However, while their timeframe is clearly defined, the rationale behind their journal selection is more ambiguous, described simply as "authoritative". Their article pool is comprised of six "influential", "high-impact" Applied Linguistics journals, including *Applied Linguistics*, *TESOL Quarterly*, *The Modern Language Journal*, *Language Teaching Research*, *Language Learning*, and *Studies in Second Language Acquisition*. While no one in the field would deny that these six journals are the top-tier journals in the field, there are more journals that are equally important in the field. Moreover, particularly within the topic of language learning motivation, journals such as *System* publish a large number of papers related to language learning motivation, so selecting journals for systematic review just based on the prestige of the journals may bias the outcomes of the review.

In another systematic review on L2 motivation research, Boo et al.'s (2015) study focused on the publications between 2005 and 2014. In the review, they classified the journals included in their data into five groups, based on the profiles of the journals (Table 3). For example, journals such as *Applied Linguistics*, *Language Learning*, and *Modern Language Journal* are labelled as Group 1, since these journals have the highest profiles in the field. Journals in Groups 3 and Group 4 are ones that reflect more localized research focus, such as *RELC Journal* which is a journal published by the South East Asian Ministers of Education Organization (SEAMEO) and *JALT Journal* published by Japan Association for Language Teaching (JALT). In this manner, Boo et al. (2015) provided more

detailed descriptions about the profiles and the statuses of journals included in the review, which makes the research process and the rationales behind data collection clearer.

Table 3. Classifications of journals included in Boo et al.'s (2015) review (adapted from Boo et al., 2015)

Group	Examples	Description
Group 1	*Applied Linguistics, Language Learning, Modern Language Journal, Studies in Second Language Acquisition, TESOL Quarterly*	The five journals that traditionally have the highest profile and generally the highest impact factors in the field of SLA
Group 2	*System, Language Teaching Research*	International, peer-reviewed Applied Linguistics research journals published by established publishers or associations
Group 3	*Canadian Journal of Applied Linguistics, RELC Journal*	Applied linguistics journals published in English-speaking countries but with a more localized focus
Group 4	*JALT Journal*	Applied linguistics journals published in non-English speaking countries
Group 5	*Journal of Language and Social Psychology*	Journals that are traditionally not associated with the field of Applied Linguistics

Boo et al.'s (2015) review tried to include studies in broader research contexts, including ones conducted in non-English speaking countries. Still, however, they only included internationally disseminated publications that are written in English, and the journals' profiles were assessed based on their global recognition, or "prestige" of the journals. It raises an issue that, since non-international journals are often not indexed in global journal databases, the profiles of these journals cannot be assessed based on the global rankings and such.

According to the SJR Journal Ranking data, the only journals indexed in Linguistics and Language and published in Japan are *CALL-EJ* (Q1), *JALT CALL Journal* (Q1), *SiSAL Journal* (Q2), *JALT Journal* (Q3), and *English Linguistics* (Q4), and no other journals are indexed in SJR's global journal database (i.e., five out of 1,347 journals). Given this background, there was a concern that a search limited to international databases alone would not be sufficient to accurately review the characteristics of the Japanese research context. Therefore, in addition to international databases such as ERIC and LLBA, the domestic Japan-

ese scholarly database CiNii was included in the search. Moreover, keywords translated into Japanese were also used to search for literature within the domestic database.

Limiting the publication period is also a common method in literature searches and can be considered a major approach in the field, adopted by many studies including Boo et al. (2015) and Mahmoodi and Yousefi (2022). This is particularly relevant for L2 motivation research, which has seen remarkable theoretical development since the 2000s. During this time, numerous theories and perspectives have been proposed and advanced, including Dörnyei's (2009) L2 Motivational Self System, Complex Dynamic Systems Theory (e.g., Dörnyei et al., 2015; Sampson & Pinner, 2020), and the rise of qualitative research. Therefore, conducting a systematic review with a restricted timeframe is justifiable, as it allows for a focused analysis of the transformation in research trends during this dynamic period. Accordingly, Aoyama (2026) reviewed research from a period similar to that covered by Boo et al. (2015) to enable a direct comparison of trends.

On the other hand, this approach has potential drawbacks. L2 motivation research has a history spanning over 60 years (Al-Hoorie, 2017) and has evolved by incorporating a wide range of theories. There are many foundational theories that have significantly contributed to its development, such as research from the social-psychological period in the 1950s (e.g., Gardner & Lambert, 1959) and major cognitive theories like Self-Determination Theory (Deci & Ryan, 1985). It is crucial to be mindful that limiting a review to recent publications will necessarily exclude these earlier studies. For this reason, when conducting a systematic review with a restricted timeframe, the review's objectives must be clearly defined and the rationale for the chosen period must be thoroughly explained.

3. Course of action

3.1 Solution to language choice bias

In response to dilemmas 1 and 2, the concerns related to the selection of publication language in systematic reviews, researchers in the field are increasingly recognizing the need to move beyond an English-only focus. Drawing on recent discussions, it is crucial that systematic reviews in Applied Linguistics incorporate studies published in languages other than English. As Chong (2025) argues, the nature of the field, with its primary focus on the highly contextualized use of languages, provides a strong encouragement for researchers to include publications from diverse linguistic contexts. This decision should be informed by the specific topic of the review and the linguistic diversity within the research team. This

aligns with the broader methodological principle articulated by Hampson and McKinley (2024), who state that "systematic reviewers need to be able to include studies of a range of backgrounds while finding holistic ways of evaluating those studies".

An English-only approach risks creating a significant disconnect between the field's explicit goal of promoting multilingualism and its actual research practices. A study of Applied Linguistics journals by Warren and Sato (2024) found that while many journals promote multilingualism in their scope, they often default to an English-only publication policy, thereby creating an incongruity between their stated commitments and their practices. This can inadvertently reinforce biases like native speakerism and marginalize valuable scholarly contributions from L2-English academics. Therefore, including research from a wider range of languages is not merely a matter of comprehensive data collection; it is an essential step toward building a more equitable, globally representative, and methodologically robust understanding of language-related phenomena.

An example of a systematic review that incorporates non-English publications is the study by Zhou et al. (2024) on cross-cultural communicative competence (ICC) in English as a Foreign Language (EFL) teaching in China. The research systematically reviews the challenges and methods related to implementing ICC skills in undergraduate English instruction within the Chinese context. A significant aspect of this review's methodology is its deliberate inclusion of peer-reviewed research articles published in both English and Chinese. This inclusive approach to data collection allows for a more comprehensive understanding of the topic, capturing insights from literature that might be missed in a review limited to English-language sources. By including Chinese-language studies, the authors ensure a deeper and more authentic analysis of the educational landscape, teacher perceptions, and assessment practices specific to China. This makes the study an excellent case for how including non-English publications in a systematic review can mitigate language bias and provide a more complete and contextually relevant synthesis of evidence, especially when the research focus is on a specific non-Anglophone region.

Additionally, some studies explicitly cite the exclusion of non-English publications as a limitation of their systematic reviews. Weng et al. (2024), for instance, discuss their English-only approach as a limitation of their study, referring to the "Tower of Babel bias" (Grégoire et al., 1995). Making efforts to include and highlight country-specific research enriches the global discourse, allowing for a more robust and more detailed picture of Applied Linguistics scholarship worldwide and ensuring that valuable local insights inform broader theoretical discussions and pedagogical practices. My own study grounds on the idea of reviewing studies on L2 motivation that have been conducted in a specific research context, by

including locally published studies in the article pool, and it shares the awareness of the issues.

Aoyama's (2026) review was designed to include domestic Japanese journals, responding to these limitations in limiting the publication language to English only. Since L2 motivation is a concept that strongly relates to learning context and backgtounds, the review specifically aimed to gather findings from research conducted within the Japanese research context. Thus, the inclusion of papers published in Japanese was expected to provide richer insights into more locally contextualized discussions on L2 motivation in the Japanese educational settings.

To include papers written in Japanese, I first explored the editorial policies and submission guidelines of Applied Linguistics-related journals within Japan. For instance, the publication language of Annual Review of English Language Education in Japan issued by the Japan Society of English Language Education is English for research articles, but submissions in Japanese are accepted for the pedagogical articles section. This implies that research articles are published in English even in domestic journals, in a similar manner as international journals, but in terms of papers with more pedagogical focus and practical approaches interconnected more strongly with classroom teaching are likely to be written and published in Japanese. Based on this, the database search was performed on both international and Japanese databases.

However, considering the practical challenge of describing all information in detail due to word count limitations in academic publications, Aoyama (2026) did not provide exhaustive details of the non-English data or direct access to the original sources. Possible alteranative approaches to show such detailed information effectively will be discussed in Section 5.2.

3.2 Inclusion of Japanese domestic journals

To address the third dilemma, Aoyama (2026) included both international journals and Japanese domestic journals in the field of Applied Linguistics and language education. The study utlized two major Japanese scholary databases, CiNii (organized by National Institute of Information: https://cir.nii.ac.jp/) and J-STAGE (organized by Japan Science and Technology Agency: https://www.jstage.jst.go.jp/). The rationale behind the inclusion of local scholarly databases is that vast majority of academic publications within Japan are included in either of these databases. For instance, J-STAGE contains over 4000 journals that include 5 million scholarly articles in its database, with direct online access to most of the articles. In addition, CiNii covers institutional bulletins called *Kiyou*, and it provides information on the libraries and institutions where the articles and books are archived if the publications are available in a print-only format. As a result,

journals locally published in Japan and not archived in international scholarly databases (e.g., ERIC, PsychINFO) were covered in the review, and the following types of domestically published papers and research reports were included in the article pool:

- English language education/Applied Linguistics journals which publish articles written in both Japanese and English (e.g., *Annual Review of English Language Education in Japan, JACET Journal, Language Education & Technology*).
- Institutional bulletin papers which publish non-refereed working papers/practical reports written in Japanese, as well as refereed research articles (e.g., *XX University Journal of Educational Research and Practice*).

As a result of including Japanese domestic journals, a greater number of publications from Japan met the inclusion criteria compared to Boo et al.'s (2015) international review. Thus, the outcome of the initial data search and screening in Aoyama's (2026) review suggests that the focus and coverage of databases can significantly affect the number of papers researchers obtain in systematic reviews, especially when the review has a specific focus on a particular research context.

In L2-related research area, there is a growing interest in research produced in diverse research context and published in non-English journals. The journal *Language Teaching* has a dedicated section entitled *A Country in Focus* which provides a platform for critical surveys of research that has been published within a specific nation. The importance of such initiative lies in its goal of bringing visibility to studies that are inaccessible to the international research community due to being published in local journals or languages other than English. The approach taken in my study serves as one possible way to support this initiative.

4. Alternative courses of action

In this section, drawing on the ongoing discussions regarding systematic review, open research, and the advancement in research tools, I will discuss potential alternative approaches that researchers can take when conducting systematic reviews in the future.

4.1 Further narrowing down the target journals

As discussed in the previous sections, Aoyama's (2026) review included both international and Japanese domestic journals. In this systematic review, the process of analyzing and attempting to balance these two distinct bodies of lit-

erature made the inherent difficulties of such approach. This experience suggests possible alternative solutions. It may be possible to create a synthesis that reflects findings from a wider range of research contexts by first conducting multiple small-scale, focused, and comprehensive systematic reviews, each with a clearly defined objective. These individual reviews could then be integrated through a bottom-up approach, a process known as a "second-order synthesis", an approach exemplified by Plonsky's (2023) study in SLA, which synthesized 308 systematic reviews to examine sampling and generalizability.

Context-specific reviews, such as the one introduced in this chapter, offer significant advantages to the field. They make it possible to synthesize research written in non-English and provide a perspective that is thoroughly informed by the background of the specific research context. In line with this approach, several systematic reviews have examined English as a Foreign Language (EFL) and wider English language education research conducted specifically within Japan, frequently analyzing articles from key domestic journals. These reviews aim to understand the distinct characteristics and trends in Japanese research, which may diverge from global patterns or appear in publications less visible in international databases.

Examples include Mizumoto et al. (2014), who reviewed articles in the *Annual Review of English Language Education in Japan* (*ARELE*) from 1990 to 2013, identifying themes, methods, research outcomes, and conducting a meta-analysis of effect sizes. Aoyama et al. (2018) performed a systematic review centred on quantitative reporting practices within *ARELE*, evaluating them against international guidelines. Honda et al. (2018) offered a retrospective systematic review of articles in the *KATE Journal* (1985–2017), coding methodological features and tracking thematic changes over time. Additionally, Aoyama (2022) carried out a preliminary review of motivation research specifically in *ARELE* to grasp trends in this area within the Japanese context. More recently, Hiromori et al. (2024) systematically reviewed motivation studies across three major domestic journals (*ARELE*, *JACET Journal*, and *JALT Journal*), notably including Japanese-language publications and focusing on extracting educational implications. Similarly, Aoyama (2022) reviewed L2 motivation study in *ARELE* journal.

The advantage of concentrating on a single research context like Japan is the potential to explore local knowledge and context-specific findings often missed by international reviews. This focused approach facilitates a detailed understanding of research conducted domestically, highlighting patterns in theoretical frameworks (e.g., the frequent application of Self-Determination Theory in some domestic motivation research), preferred methodologies, common participant groups (e.g., the significant representation of university students), and typical reporting standards in local journals. Context-specific reviews are also better

positioned to identify research addressing local educational policies or societal changes, as suggested by Hiromori et al.'s (2024) focus on domestic journals addressing the local context, and to derive practical implications suited to the particular teaching and learning environment. Analysing domestic journals thus yields insights that enrich the understanding of the research field within that specific national context (see Chapter 1 of this book).

Furthermore, in addition to systematic reviews with a "country-in-focus" perspective, there are also review studies that target a single journal to understand research trends on a more specific topic. For instance, Zhang et al. (2022) conducted a systematic review of L2 motivation studies published in the journal *System*, focusing specifically on research conducted in school settings. They discussed the significance of this topic and the implications for future research.

This suggests that accumulating focused systematic reviews, such as those with a narrow journal scope, and then conducting a broader, overarching discussion through a second-order synthesis could have a significant impact on the field of Applied Linguistics. This methodological approach, therefore, warrants further consideration.

4.2 Effectively analyze non-English studies

A central point of discussion in this chapter is how to handle research published in languages other than English. When the objective of a systematic review is to interpret the overall landscape of a research field, analyzing studies written in languages other than English presents significant challenges. For this reason, limiting the analysis to English-language publications has, by necessity, become the mainstream approach. However, what methods can be employed to overcome these challenges and include local studies written in non-English languages in systematic reviews?

One potential solution is the effective utilization of generative AI. The rapidly changing technological development in this area has the potential to contribute significantly to future systematic reviews, as recent advancements have made it considerably easier to comprehend materials written in unfamiliar languages. In their review of research in the Japanese context, Terasawa and Segawa (2025) also note that recent progress in machine translation has substantially lowered the language barriers that have hindered previous review studies, making it more feasible to include non-English publications. While fully automating the literature collection and analysis processes with generative AI is not yet practical due to concerns about coding consistency and reliability, which necessitate human involvement (Clark et al., 2025; also see Chapter 17 of this book), it can certainly be utilized to

better embrace the multilingual and multicultural nature of the field of Applied Linguistics.

Another potential strategy is for researchers from diverse backgrounds to collaborate on the analysis of studies conducted in different contexts. Stern and Kleijnen (2020) suggest that instead of costly full translations, a more economical solution is for reviewers to work with individuals who can read the language to identify and extract the necessary information. Taking the example of L2 motivation, research contexts span numerous countries across the globe, including many in Asia (Boo et al., 2015). By conducting systematic reviews that incorporate local research within such global research topics, it becomes possible to gather and synthesize a much broader range of insights on the subject.

5. Conclusion and future directions

In this chapter, drawing on my own example of a systematic review, practical questions, dilemmas, and solutions that have emerged during the research process are discussed. The research questions of the original study (Aoyama, 2026) are specifically designed to address the identified gaps by exploring the complex and dynamic changes in English language learners' motivation in Japan, and its associations with L2 motivation constructs as well as sociocultural factors within the specific educational context.

Considering the study's nature of putting emphasis on the inclusion of studies conducted in a particular research setting, the discussions might not be directly applicable to other systematic reviews in the field of Applied Linguistics. However, I believe that the issues raized in this chapter need to be discussed more broadly among the researcher community in the field of Applied Linguistics, in order for the field to effectively summarize and synthesize the accumulation of research to move forward.

Particularly, with diversifying and globalizing research trends involving the presence of a more multilingual research in the field, it is essential that we further consider how the nature of the field can be treated in systematic reviews. Revisiting the practices of systematic reviews is another essential task for researchers in Applied Linguistics. Since the 2000s, following early influential works in the field (e.g., Norris & Ortega, 2000; Masgoret & Gardner, 2003), a substantial number of systematic reviews and meta-analyses have been produced. It is now necessary to analyze the methods employed in these past syntheses in order to discuss and refine future approaches for conducting systematic reviews in Applied Linguistics. In this regard, Chong's (2025) SMART (Synthesis Methods and Reporting

Tool) is an excellent initiative that deserves to be more widely known and utilized within the field.

Furthermore, it would also be important that we capture how researchers in the field perceive the issues discussed in this chapter. Particularly, from the viewpoint of language choice and focus on local research contexts, it is crucial that we listen to the voices of researchers from different contexts to explore the needs and to better understand their perceptions about systematic reviews in the field.

6. Questions for discussion

- If you were to conduct a systematic review on a topic of your choice in Applied Linguistics, what article search strategies and inclusion/exclusion criteria would you establish? How would you address the issue of publication language?
- Based on the nature of your chosen research topic, how well could you explain and justify the rationale behind the approach you outlined above?
- In your own area of research, what is your understanding of the geographical distribution of the studies being conducted? If you were to perform a systematic review in this area, are there studies from specific regions or works in particular languages that you believe must be included to ensure a comprehensive analysis?

Further reading

doi Chong, S.W. (2025). Synthesis methods and reporting tool (SMART) for research syntheses in Applied Linguistics. *Research Synthesis in Applied Linguistics*, *1*(1), 17–38.

Chong (2025) presents the Synthesis Methods and Reporting Tool (SMART), a comprehensive guide for conducting and reporting high-quality research syntheses in applied linguistics. This paper strongly reinforces the importance of incorporating diversity and reflexivity, serving as an excellent methodological guideline for researchers aiming to capture a broader range of research, including studies published in different languages and contexts.

doi Hampson, T., & McKinley, J. (2024). Systematic review, systematic bias? An example from EMI research. *Language Teaching*, 1–9.

Hampton and McKinley (2024) offers a critical perspective on systematic reviews, arguing that their positivist methodology and narrow focus can pose a risk by presenting an incomplete picture of a research field. This paper serves as an insightful reminder that as researchers, we must be mindful of the paradigm that informs a systematic review and use its findings with explicit criticality, rather than treating them as definitive conclusions for a broader research area.

Terasawa, T., & Segawa, H. (2025). A critical review of L2 teaching and learning research in Japan (2019–2023). *Language Teaching*, *58*(3), 329–358.

Serving as a prime example of a country-in-focus approach, Terasawa and Segawa's (2025) paper offers a critical review of L2 teaching and learning research in Japan from 2019 to 2023. It synthesizes findings from papers published in both English and Japanese, highlighting key developments and challenges relevant to L2 teaching in Japan for a global readership.

AI use declaration statement

During the preparation of this work, the author used Google Gemini for writing assistance and proofreading. As author, I confirm that after using the tool, I reviewed and edited the content as needed and take full responsibility for the content of the publication, in alignment with COPE and the publisher's AI policy.

References

Al-Hoorie, A. H. (2017). Sixty years of language motivation research: Looking back and looking forward. *Sage Open*, *7*, 1–11.

Al-Hoorie, A. H., Oga-Baldwin, W. L. Q., & Hiver, P. (2025). Self-determination mini-theories in second language learning: A systematic review of three decades of research. *Language Teaching Research*, *29*, 1603–1638.

Aoyama, T. (2022). Language learning motivation research in ARELE: A review of research trends and characteristics. *Shinshu University Journal of Educational Research and Practice*, *16*, 1–16.

Aoyama, T. (2026). Exploring Japanese students' journeys of learning English: A mixed-methods retrospective analysis (Unpublished doctoral dissertation). University of Warwick.

Aoyama, T., Hoshika, M., & Honda, K. (2018). Reporting practices in L2 quantitative research: A systematic review of ARELE 13–28. *Annual Review of English Language Education in Japan*, *29*, 241–256.

Apple, M. T., Da Silva, D., & Fellner, T. (Eds.). (2017). *L2 selves and motivations in Asian contexts*. Multilingual Matters.

Aryadoust, V., Soo, Y. X. N., & Zhai, J. (2024). Exploring the state of research on motivation in second language learning: A review and a reliability generalization meta-analysis. *International Review of Applied Linguistics*, *62*, 1093–1126.

Boo, Z., Dörnyei, Z., & Ryan, S. (2015). L2 motivation research 2005–2014: Understanding a publication surge and a changing landscape. *System*, *55*, 145–157.

Boutron, I., Page, M. J., Higgins, J. P. T., Altman, D. G., Lundh, A., & Hróbjartsson, A. (2019). Considering bias and conflicts of interest among the included studies. In J. P. T. Higgins, J. Thomas, J. Chandler, M. Cumpston, T. Li, M. J. Page, & V. A. Welch (Eds.), *Cochrane Handbook for Systematic Reviews of Interventions* (2nd ed., pp. 177–204). Wiley-Blackwell.

Chong, S. W. (2025). Synthesis Methods and Reporting Tool (SMART) for research syntheses in Applied Linguistics. *Research Synthesis in Applied Linguistics*, *1*, 17–38.

Clark, J., Barton, B., Albarqouni, L., Byambasuren, O., Jowsey, T., Keogh, J., Liang, T., Moro, C., O'Neill, H., & Jones, M. (2025). Generative artificial intelligence use in evidence synthesis: A systematic review. *Research Synthesis Methods*, *16*(4), 601–619.

Csizér, K., & Magid, M. (Eds.). (2014). *The impact of self-concept on language learning*. Multilingual Matters.

Deci, E. L., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior*. Plenum.

Dörnyei, Z. (2009). The L2 motivational self system. In Z. Dörnyei & E. Ushioda (Eds.), *Motivation, language identity and the L2 selves* (pp. 9–42). Multilingual Matters.

Dörnyei, Z., & Ushioda, E. (Eds.). (2009). *Motivation, language identity and the L2 self*. Multilingual Matters.

Dörnyei, Z., MacIntyre, P. D., & Henry, A. (2015). *Motivational dynamics in language learning*. Multilingual Matters.

Gardner, R. C., & Lambert, W. E. (1959). Motivational variables in second language acquisition. *Canadian Journal of Psychology*, *13*, 266–272.

Grégoire, G., Derderian, F., & Le Lorier, J. (1995). Selecting the language of the publications included in a meta-analysis: Is there a tower of babel bias? *Journal of Clinical Epidemiology*, *48*, 159–163.

Güleçoğlu, G. Y., & Öztürk, G. (2021). Research perspectives on English as a Foreign Language (EFL) learning motivation in Turkish context: A systematic review of studies between 2010 and 2021. *Language Teaching and Educational Research (LATER)*, *4*, 161–180.

Hampson, T., & McKinley, J. (2024). Systematic review, systematic bias? An example from EMI research. *Language Teaching*, 1–9.

Hiromori, T., Okunuki, A., Kashimura, Y., Tterao, K., Kanemoto, S., Suzuki, H., Nakamura, H., & Izumisawa, M. (2024). Motivation in English language learning: A systematic review of three major domestic journals over the past 20 years. *Annual Review of English Language Education in Japan*, *35*, 33–48.

Honda, K., Hoshika, M., Aoyama, T., Someya, F., & Yamamoto, T. (2018). A systematic review of articles in KATE 1–31: The changing trends in the field of English education. *KATE Journal*, *32*, 85–98.

Jüni, P., Holenstein, F., Sterne, J., Bartlett, C., & Egger, M. (2002). Direction and impact of language bias in meta-analyses of controlled trials: empirical study. *International Journal of Epidemiology*, *31*, 115–123.

MacIntyre, P. D., Gregersen, T., & Mercer, S. (Eds.). (2016). *Positive psychology in SLA*. Multilingual Matters.

Masgoret, A. M., & Gardner, R. C. (2003). Attitudes, motivation, and second language learning: A meta-analysis of studies conducted by Gardner and associates. *Language Learning*, *53*, 167–210.

Mizumoto, A., Urano, K., & Maeda, H. (2014). A systematic review of published articles in ARELE 1–24: Focusing on their themes, methods, and outcomes. *Annual Review of English Language Education in Japan*, *25*, 33–48.

Mahmoodi, M. H., & Yousefi, M. (2022). Second language motivation research 2010–2019: a synthetic exploration. *The Language Learning Journal*, *50*(3), 273–296.

Morrison, A., Polisena, J., Husereau, D., Moulton, K., Clark, M., Fiander, M., Mierzwinski-Urban, M., Clifford, T., Hutton, B., & Rabb, D. (2012). The effect of English-language restriction on systematic review-based meta-analyses: A systematic review of empirical studies. *International Journal of Technology Assessment in Health Care*, *28*, 138–144.

Norris, J. M., & Ortega, L. (2000). Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis. *Language Learning*, *50*, 417–528.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Aki, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *PLOS Medicine*, *18*(3), e1003583.

Plonsky, L. (2023). Sampling and generalizability in Lx research: A second-order synthesis. *Languages*, *8*, 75.

Sampson, R., & Pinner, R. (2020). *Complexity perspectives on researching language learner and teacher psychology*. Multilingual Matters.

Shurovi, M., Yahaya, M. F., Hajimia, H., & Hasan, M. K. (2025). A systematic literature review of ESL/EFL learning strategies and learner motivation. *International Journal of Learning, Teaching and Educational Research*, *24*, 252–277.

Stern, C., & Kleijnen, J. (2020). Language bias in systematic reviews: You only get out what you put in. *JBI Evidence Synthesis*, *18*, 1818–1819.

Suri, H., & Clarke, D. (2009). Advancements in research synthesis methods: From a methodologically inclusive perspective. *Review of Educational Research*, *79*, 395–430.

Terasawa, T., & Segawa, H. (2025). A critical review of L2 teaching and learning research in Japan (2019–2023). *Language Teaching*, *58*(3), 329–358.

Warren, L. K., & Sato, M. (2024). Multilingualism and native speakerism in academic journals' language policies: Exploring a potential power of applied linguistics journals in promoting equitable publishing practices. *Studies in Second Language Acquisition*, *46*, 921–932.

Weng, Y., Schmidt, M., Huang, W., & Hao, Y. (2024). The effectiveness of immersive learning technologies in K-12 English as second language learning: A systematic review. *ReCALL*, *36*, 210–229.

Wu, X. (2022). Motivation in second language acquisition: A bibliometric analysis between 2000 and 2021. *Frontiers in Psychology*, *13*, 1032316.

Zhang, F., Wang, J., & Hennebry-Leung, M. (2022). Research on language learning motivation in school settings in System. *System*, *107*, 102817.

Zhou, R., Samad, A., & Perinpasingam, T. (2024). A systematic review of cross-cultural communicative competence in EFL teaching: Insights from China. *Humanities & Social Sciences Communications*, *11*, Article 1750.

CHAPTER 15

Sing a song of synthesis

Harmony and discordances in preparing a systematic review on the effects of songs in second language education

Hamish Chalmers & Catherine Hamilton
University of Oxford

One of the key expectations of doctoral research is that it must contribute original knowledge. Systematic reviews can demonstrate authoritatively where gaps to be filled exist. This chapter describes a partnership between a doctoral researcher and her supervisor (the authors) as they sought to comprehensively map and evaluate the field of existing intervention research on the effects of singing songs with young language learners on their substantive linguistic outcomes. It will describe the thought processes behind conducting a methodologically robust and ethically sound review, outline the methodological challenges and solutions encountered while locating, classifying, appraising and synthesising a heterogenous body of research, and recount their experiences when submitting the review for publication, where careful communication of their contra-orthodox findings was required to satisfy a review panel of established researchers.

Keywords: systematic review, songs, music, language learning, methodology

1. Background

This chapter explores the process of preparing a systematic review as part of a doctoral student's research journey. In particular, we describe how doctoral student (CH), and her supervisors (HC, and Victoria Murphy), navigated the conception, design, execution and publishing of a systematic review exploring the substantive linguistic effects of using songs for teaching young language learners, which formed the first phase of the doctoral project. Prior to embarking on her project, CH had identified a mismatch between practitioners' strongly held beliefs about the effectiveness of using songs for teaching young language learners and

https://doi.org/10.1075/rmal.17.15cha
Available under the CC BY-NC-ND 4.0 license.© 2026 John Benjamins Publishing Company

the apparently much less strong evidence base to support that position (Hamilton & Murphy, 2023). Using songs in the languages classroom is a popular approach with teachers of foreign languages in England (ibid.) and internationally (Garton et al., 2011; Harris & O'Leary, 2009). Surprisingly, according to previous literature reviews (Davis, 2017; Sposet, 2008; Werner, 2020) there appeared to be scant evidence from robust research to underpin teachers' pedagogical decisions about the judicious use of songs for the purpose of achieving linguistic outcomes. Thus, there appeared to be a clear case for robustly assessing whether the centuries of 'folk pedagogy' (Bruner, 1996) that promoted a perceived benefit of using songs to achieve linguistic outcomes could be stood up by research evidence.

The reviews by Davis (2017), Sposet (2008), and Werner (2020) had not been conducted in a wholly transparent or systematic way, and, importantly, had not included systematic appraisal of the methodological quality of the assembled primary research. Therefore, despite their general conclusions that 'songs work', it was impossible to ascertain the strength of the evidence upon which that claim was based. The systematic review that ensued, therefore, was conceived to address two aims. First, we sought to ascertain the extent and nature of intervention research investigating the effects on linguistic outcomes of using songs with school-aged learners of second or foreign languages. Second, we aimed to assess what could be reliably concluded from that body of evidence, through a narrative synthesis of the research and an accountable approach to appraising methodological quality.

Eligible studies were those conducted in preschool, primary or secondary schools, in which songs or nursery rhymes were delivered as whole-class interventions, and which measured substantive linguistic outcomes. These outcomes were defined as any measure of vocabulary, grammar, speaking, listening, reading, or writing. CH's prior (non-systematic) reading of the relevant literature had exposed very few primary studies in this area, and an apparent paucity of robust experimental designs in particular. We therefore did not wish to bias the review by excluding potentially relevant work based on research design, language of publication or publication status (i.e. theses or peer-reviewed articles).

Following a search of relevant English, Spanish, French and German language databases, 1862 abstracts and 89 full texts were screened, resulting in 60 eligible intervention studies published in 23 countries between 1979 and 2021. Given the heterogeneity of participants, methods and outcome measures in the included works, we conducted a narrative synthesis of the findings, rather than a meta-analysis. Most of the assembled studies made positive causal claims about the relationship between using songs and linguistic outcomes. However, our formal assessment of the methodological quality of that body of evidence, conducted using the Mixed Methods Appraisal Tool (MMAT; Hong et al., 2018), suggests

that it is impossible to draw any firm conclusions about the veracity of those claims. The cumulative weight of evidence, therefore, is limited in its capacity to help teachers or policymakers make evidence-informed decisions about related practice. Our completed review (Hamilton, Schulz, Chalmers, & Murphy, 2024) makes the case for conducting further robustly designed and reported intervention research to better inform our collective understanding of the linguistic effects of using songs or nursery rhymes for teaching languages to young learners.

The following section outlines the principles and motivation for preparing systematic reviews as part of a programme of doctoral research, and the methodological and ethical dilemmas we considered in doing so.

2. Methodological and ethical dilemmas

A systematic review is expected to provide an authoritative and comprehensive synthesis of the existing body of knowledge on a given topic. As the Evidence for Policy and Practice Information and Co-ordinating Centre (EPPI Centre) at the UCL Institute of Education states:

> The key features of a systematic review or systematic research synthesis are that:
>
> – explicit and transparent methods are used
> – it is a piece of research following a standard set of stages
> – it is accountable, replicable and updatable
>
> … Systematic reviews aim to find as much as possible of the research relevant to the particular research questions, and use explicit methods to *identify what can reliably be said* on the basis of these studies. (EPPI Centre, 2019a, n.p., emphasis added)

The emphasised text here underscores a crucial point: it is not sufficient to merely assemble a body of literature pertaining to a specific research question. Rather, a systematic review must also provide a synthesis of the evidence that can inform relevant stakeholders of the reliability of that research, enabling readers to discern what conclusions can be drawn with confidence. As such, methodological and ethical considerations are inextricably linked throughout the review process.

If the aim of a systematic review is to determine what can be reliably stated about an area of inquiry, reviewers must assume an ethical responsibility to ensure that any claims made are defensible in the light of the methods employed. These methods include the strategies for identifying, selecting, analysing, and interpreting relevant literature. This obligation arises not only from a scholarly commitment to rigour and impartiality, but also from a duty to the end-users of systematic reviews who rely on reviews to guide their decision-making.

As Gough, Oliver, and Thomas (2012) observe:

> [F]or policy makers, practitioners, and people in general making personal decisions, engaging in mountains of individual research reports, even if they could find them, would be an impossible task. Instead, they rely on researchers to keep abreast of the growing literature, reviewing it and making it available in a more digestible form. (pp. 1–2)

Gough, Oliver and Thomas thus highlight the ethical imperative to conduct methodologically robust systematic reviews. This entails clearly defining the scope of the review so that the conclusions are understood in context, articulating transparent inclusion and exclusion criteria so that readers understand how the literature was selected, and evaluating the methodological quality of the included studies. Methodological evaluation ensures that conclusions are informed not just by the findings reported but also by an assessment of how well executed the research was, and thus how trustworthy those findings are.

Before undertaking a systematic review, it is also necessary to determine whether there is sufficient justification — or warrant — for conducting the review at all. This includes ensuring that the review addresses a meaningful gap in knowledge and that it will provide value to intended audiences. These considerations are encapsulated in the notion of 'avoidable research waste', which guided our thinking in preparing this review.

2.1 Research waste and the ethics of conducting a systematic review

Isaacs and Chalmers (2025) articulate the principles behind the notion of 'avoidable research waste' in Applied Linguistics. They drew on a concept introduced into the literature on ethical research in the field of healthcare (Chalmers & Glasziou, 2009; Macleod et al., 2014), and which expanded the discussion of ethics beyond the standard concerns (protecting personal data, ensuring participants come to no harm during the course of research, and so on) into a larger conceptualisation of ethics: primarily the ethical considerations one must take into account when deciding what to research and how. They proposed five potential sources of avoidable research waste that must be accounted for when planning and conducting any new research. These are:

- The relevance of the research question (i.e. are the research questions meaningful and relevant to the stakeholders' interests and concerns?)
- Necessity of the research and appropriateness of the design (i.e. is there genuine uncertainty around the potential answer to the research question — more

simply, does an answer to the research question already exist — and if so, is the proposed research design well-suited to addressing the question robustly?)
- Research regulations and management (i.e. are regulatory and management processes appropriate and consistent?)
- Publication and dissemination of results (i.e. on completion of the research, are the findings properly disseminated to all relevant stakeholders?)
- Quality of reporting (i.e. is the research reported in a way that ensures readers can fully understand how the research was conducted and thus be in a position to critically evaluate its findings?).

(After Isaacs & Chalmers, 2025; Macleod et al., 2014)

If these five categories have not been appropriately considered when embarking on new research (primary or secondary), and the answers to their related questions cannot be answered in the affirmative, then researchers risk conducting wasteful research. Wasteful research is unethical research. Research always costs something, whether that is in fiscal terms (the grants that fund research, which ultimately often come from the public purse), or in terms of time, effort and opportunity on the part of those involved in the research. It is, therefore, unethical to conduct research that squanders these limited resources. It was our intention when embarking on this project to adhere as closely as possible to a framework that minimised waste in the process and thus maximised the contribution our review stood to make.

For the purpose of conducting the systematic review discussed here, we had most control over the first, second, and last categories in the taxonomy of waste outlined above. That is, we were in a position to ensure that the question driving the review was meaningful and relevant; we were able to make an informed decision about how necessary our review was likely to be, and to decide carefully on a design that would robustly address the question we asked; and we could ensure that it was reported such that readers would understand what we did and how we did it, all of which we will elaborate on in Section 3, 'Course of action'.

2.1.1 *The relevance of the research question*

Applied research is always conducted with an end user in mind. For example, in Applied Linguistics research there is an expectation, *inter alia*, that research will help teachers and other educators to understand how languages are most effectively taught and learned. There is an expectation that addressing questions of efficacy in language teaching will inform, and hopefully improve, decision making by the relevant policy makers and practitioners. Applied Linguistics research that demonstrates an understanding of what these groups see as priorities around their practice stands to produce better, more impactful, less wasteful (and there-

fore more ethical) research. Developing a good understanding of the priorities of the intended audience helps to avoid doing research that is likely to go to waste because it addresses questions that no one is interested in. It helps researchers to understand what outcome measures these groups value and take this into account when planning research. Ultimately, research informed by user priorities is more likely to effect positive change. Researchers, therefore, should aim to understand what priorities and uncertainties exist among their intended audiences and address research questions that matter to them. This is just as important for new secondary research (or research synthesis) as it is for new primary research.

HC and CH have both been languages teachers and thus have at least some first-hand knowledge of the kinds of things languages teachers consider important. HC was a primary school teacher in the international sector, whose pupils were, in the majority, EAL learners. He regularly sang English language songs with his pupils. CH was a secondary school teacher of French and English, then founded an organisation called Babel Babies, where preschoolers explored the languages of the world through multilingual songs and stories. We both felt that using songs in this way was a good use of our time. In addition to being fun activities that our young learners seemed to enjoy, we assumed that there was a more utilitarian educational purpose at play. We chose to sing songs with our learners in part because we felt that songs were an effective medium by which to present and rehearse foreign language topics, vocabulary, and grammar. In short, we felt that using songs helped children to learn new languages.

When CH started the research agenda that culminated in her doctoral project, she surveyed other teachers to find out how and why they use songs with young learners (Hamilton & Murphy, 2023). Through questionnaires and interviews with early education practitioners, she found that they held strong beliefs about the effectiveness of songs "for teaching content, establishing classroom routines, and supporting pupils' linguistic and social development" (p. 1488). She also began to look carefully at research that might underpin those beliefs. In her non-systematic engagement with the literature, the research she found tended to be poorly theorised, questionably conducted, self-referential, and ultimately unconvincing. This was especially true for the idea that songs are effective in promoting linguistic development. Thus, there seemed to be a clear sense of collective uncertainty around what can be reliably concluded about the educational effects of singing songs with young language learners. The significant enthusiasm she observed among teachers and their belief that songs were not just a good way but, in some cases, a superlative way to teach languages did not seem to be supported by the research that she was aware of. Consequently, there appeared to be a strong warrant for finding out if the research she was aware of was reflective of the field in general, and whether engaging in an exhaustive search, evaluation, and synthe-

sis of the relevant evidence could reliably inform practice. This brings us to the second category of avoidable research waste: necessity and design.

2.1.2 *Necessity of the research*

In addition to understanding whether the users of research are likely to find the questions it addresses to be meaningful and relevant, researchers should ensure that addressing those questions is necessary. In essence, this means understanding whether an answer to the question already exists. Conducting new research that addresses a question to which an answer is already known, or conceivably known, is a waste of resources and, thus, ethically questionable. This is where the importance of well-conducted systemic reviews, and using existing reviews to inform decisions about new systematic reviews, comes into play.

The importance of systematic reference to existing evidence before conducting new research has long been understood. Though not the first to use the aphorism, Issac Newton remarked in 1675 that "If I have seen further it is by standing on the shoulders of Giants" (quoted in Turnbull, 1959, p. 416). More recently, Moher et al. (2010) recommend that "at a minimum, the discussion [of the findings of a new study] should be systematic and based on a comprehensive search" (p. 48). Clarke et al. (2024, p. 212) are more prescriptive: "Plans for new research should use systematic reviews of relevant existing research to provide the scientific and ethical justification for the design and conduct of new studies." This principle (applied here to new primary research) is no different for systematic reviews. The Preferred Reporting Standards for Systematic Reviews and Meta-Analyses (PRISMA) statement (Page et al., 2021, n.p.) urges authors planning new systematic reviews to "describe the current state of knowledge and its uncertainties [and] [i]f other systematic reviews addressing the same (or a largely similar) question are available, explain why the current review was considered necessary."

We located three reviews that had addressed the same or similar questions that were motivating CH's doctoral project. The oldest of these (Sposet, 2008) describes itself as a bibliography of research on the role of music in second language acquisition. It does not include information about the search and selection processes, nor what constituted eligible evidence. Fifteen of the 23 included studies asserted positive outcomes associated with using music for second language acquisition. However, the review did not include an appraisal of study quality, thus complicating critical appraisal of the overall conclusions of the review. Davis (2017) conducted what he called a 'critical review' of the evidence on using songs in young language learners' classrooms. He found nine eligible studies and concluded that there might be pedagogical value in using songs. While Davis summarised his search and selection strategy, this lacked the detail necessary to faithfully replicate it and appeared limited in scope. Like Sposet (2008), Davis

did not conduct a quality appraisal of the research he assembled. Thus, critical evaluation of his conclusions was not possible. The most recent review (Werner, 2020) was the most robustly reported. Werner stipulated that only experimental or quasi-experimental research that assessed the effects of "lyrics-based" (p. 139) instruction on different elements of language proficiency was eligible. He only searched Google Scholar and "the archives of SLA journals" (p. 143), but did not say which ones. The 28 studies that Werner located draw a mixed picture. While generally concluding that songs have a positive effect on some areas of language learning, Werner notes that songs appear to be no better than alternative ways of teaching. He, like Sposet and Davis, did not include a formal quality appraisal of the assembled research. He did, however, note methodological shortcomings, such as a lack of pre-testing of comparison groups, trivially short durations of interventions, lack of delayed posts-tests, lack of statistical controls, and failure to describe the interventions being evaluated. None of the previous reviews included reports of research that had been published in languages other than English.

From our reading of these reviews, we concluded that a new and updated review was warranted. The previous reviews did not report their methods in sufficient detail to understand them fully or to faithfully replicate them, they were limited in scope to only studies published in English, and they left questions about the level of trust we can put in their overall conclusions because they did not appraise the methodological quality of the assembled research. Aspects like quality appraisal and full and transparent reporting of search strategies are elements of good systematic review design, which brings us to the second part of this category of research waste.

2.1.3 *Appropriateness of review design*

In their taxonomy of avoidable research waste, Isaacs and Chalmers (2025) stress the importance of adopting a design appropriate for addressing the question that the study seeks to understand. If the design cannot adequately address the question, it is wasteful and unethical. CH observed this problem in some of the primary studies she came across in her non-systematic reading of the literature and from scrutinising the primary studies reported in the three reviews summarised above. In the main this was research that purported to address questions of causality using designs that are not well-suited to identifying causality: case studies of single groups of learners, with no comparison, for example. As with primary research, so with secondary.

If we are to agree with the EPPI Centre (2019a, n.p.) that systematic reviews should "find as much as possible of the research relevant to the particular research questions, and use explicit methods to identify what can reliably be said on the basis of these studies," then we must ensure that we adopt a design capable of

doing so. Robust systematic review design is (or should be) well known. Organisations like Cochrane, the EPPI Centre, the Campbell Collaboration, and the Joanna Briggs Institute, all well-regarded, longstanding producers of systematic reviews, have provided methods guidance for producing high-quality systematic reviews for decades. Nonetheless, despite this well-established history, it is unusual to find these meticulous methods adopted, or adopted in full, by systematic reviewers. Indeed, Ioannidis (2016) laments the state of the art with regards to rigour in the design of systematic reviews. He describes "massive production of unnecessary, misleading, and conflicted systematic reviews" (p. 485) which are often flawed in their "design, conduct, analysis, and reporting" (p. 501). Ioannidis was speaking in the main about systematic reviews of healthcare research, but the situation is no better in Applied Linguistics. Methodological audits of systematic reviews and meta-analyses in Applied Linguistics (Chalmers, Brown, & Koryakina, 2024; Vuogan & Li, 2024) have found serious shortcomings in the design, conduct, and reporting of much of this type of research.

We wanted to ensure our systematic review adhered as closely as possible to best practices in the field and avoided the shortcomings we had witnessed in other reviews. As we have noted, there is a great deal of guidance on systematic review design from organisations like Cochrane (Higgins et al., 2024), the Campbell Collaboration (Aloe et al., 2024), the EPPI Centre (EPPI Centre, 2019b), and the Joanna Briggs Institute (Aromataris et al., 2024). All provide information about the design elements they consider to be important to include when planning and conducting a systematic review. While there is minor variation in the specifics expounded by each of these organisations, and variation related to the type of systematic review proposed, all tend to coalesce around a common series of items that should be included in any systematic review. Conveniently, this series of items is condensed and laid out in the PRISMA reporting guideline, which brings us to the third category of avoidable research waste: quality of the reporting.

2.1.4 *Quality of the reporting (and by extension the method)*

"It must be assumed […] that any study being considered in terms of trustworthiness has been fully and clearly explained. If [the report is] not clear then the reader is justified in assuming the worst. Put simply, a poorly described study cannot be trusted" (Gorard, 2015, p. 2). A study that cannot be trusted because it has not been transparently reported is a study that is wasteful and unethical. Anyone who has prepared a systematic review will have been frustrated by the way so much primary research is reported: abstracts that merely hint at how the research was conducted and what it found, methods sections that do not describe the sample in any detail, missing statistical information, incomplete outcome data, and even a failure to describe the interventions being evaluated. This matters. Espe-

cially for research synthesists, who rely on complete and transparent reporting to make informed decisions about the eligibility of a study for their synthesis and, if eligible, have the information they need to understand it, critically evaluate it, and synthesise it with other similar studies. Research synthesists must be equally vigilant with their own reporting to avoid producing work that cannot be fully understood, and which, therefore, leaves questions about what it contributes. Both primary and secondary researchers are helped in ensuring that key information is not omitted from their reports when they adhere to industry-standard reporting guidelines.

Reporting guidelines are defined by the Equator Network (an organisation that creates, curates, and disseminates guidelines for a multitude of different research designs) as a minimum list of information needed to ensure a research study can be understood by a reader, replicated by a researcher, and used by a practitioner to make practice decisions (Equator Network, n.d.). We wanted to ensure that we adhered as closely as possible to the industry standard in our review, to avoid charges of committing the same reporting missteps that we have levelled at others. More importantly, adhering to the industry standard would maximise the possibility that our review would make a meaningful contribution to the state of knowledge, and be amenable to fair critique, replication, and extension. The industry standard for systematic reviews is PRISMA. PRISMA is a checklist of 27 items that should be addressed in the manuscript of a systematic review. It includes basic administrative items, such as identifying the report as a systematic review and providing a statement of its objectives. It also includes requirements for more detailed information about the methods and procedures. For example, full information about eligibility criteria, the search and selection strategy, outcomes for which data were sought, trustworthiness appraisal, and the rationale for the method of synthesis. One of the advantages of using PRISMA as a guideline for reporting is that it has a wash-back effect on the method. That which must be reported, must first be conducted. PRISMA was created (and is monitored) by an international group of experienced evidence synthesists and is explicitly endorsed by around 200 journals and editorial organisations (PRISMA, 2025). Using PRISMA as our guide, we could be more confident that our review would be conducted to internationally recognised standards and reported clearly and comprehensively.

2.2 Summary of ethical concerns

To summarise: we conceptualise ethics in the context of preparing systematic reviews as an obligation on the part of reviewers to produce robust and defensible syntheses that are conducted in the spirit of maximising informative potential

while ameliorating sources of avoidable research waste. To do otherwise is to risk producing syntheses that cannot reliably inform policy and practice, which compound uncertainty associated with the primary research synthesised, and which waste time, energy and resources. For our systematic review, this meant aspiring to ensure that (a) the question we addressed was meaningful to teachers, (b) the review was warranted on the basis of existing research, (c) that our methods were comprehensive and robust, and (d) our research was completely and transparently reported. Our aproach to this is explored further in Section 3.

3. Course of action

Having outlined our disposition on the methodological and ethical issues around conducting research syntheses, we turn to the specific approaches we adopted for this review, with a focus on areas that we felt deserved particular attention: exhaustive searching, and appraisal of the methodological quality of the assembled literature.

3.1 Approach to synthesis

We used a systematic review design (see Section 7 overview), with narrative synthesis. Systematic reviews as operationalised by, *inter alia*, Gough, Oliver and Thomas (2012) and Petticrew and Roberts (2006), and which are informed by following PRISMA guidelines (Page, et al. 2021), provide a strong basis for informing policy and practice. They demand meticulous attention to method and provide a commonly understood and well-regarded framework for complete and clear reporting. We elected to synthesise our findings narratively, as the diverse methodological nature of the research that met our inclusion criteria did not lend itself to statistical synthesis.

We will not rehearse the minutiae of conducting a PRISMA-guided systematic review here, as this method is well known (though not always well-adhered to), can be easily accessed, and is described in detail in the paper (Hamilton et al., 2024). However, we feel further commentary on our approach to searching for eligible literature and our method of quality appraisal is warranted, as these were two areas of concern associated with previous syntheses on the topic.

To provide context, the research questions we aimed to address were as follows:

RQ1: What is the extent and nature of intervention research investigating the substantive linguistic effects of using songs to teach second or foreign languages to young learners in formal education contexts?

RQ2: What can be reliably concluded from intervention research identified in RQ1 about the effects of using songs with young language learners on substantive foreign language learning outcomes?

By aiming to assemble, describe, and evaluate as close to the totality of existing intervention research as possible, we hoped to avoid the biases associated with selective reporting, and thus provide as complete and authoritative summary of the evidence as we were able. It is common to find systematic reviews that explicitly exclude publications in languages other than English (e.g. Nitami & Santosa, 2024; Du & Daniel, 2024; Arslan, 2020). We believe that explicit exclusion of non-English language publications risks locating a biased sample of the evidence, rather than its totality. Our review team contained speakers of French, German and Spanish as well as English. Accordingly, we constructed comparable search strategies in all these languages, and applied the search to relevant French, German and Spanish language bibiliographic databases, in addition to English language databases. While far from exhausting the possible languages of publication, this went some way to addressing linguistic bias. In addition, it is common to find English language abstracts for studies published in other languages. We committed, therefore, to recruiting the help of speakers of languages not represented in our review team to screen and extract data from any publications where an English language abstract suggested an otherwise non-English report might be eligible. In the event, we found five studies to which this applied. Whilst we cannot claim that the review is truly multilingual, it nonetheless demonstrates a way forward for addressing the long-standing monolingual English bias in research synthesis in Applied Linguistics.

In Section 2.1.3 we outlined the importance of following PRISMA guidelines. For us, this began by preparing the review protocol following the PRISMA extension for protocols (Moher et al., 2015) and registering this on IDESR.org, a systematic review protocol registry. There is no space to elaborate on the importance of prospective protocol registration here. Suffice it to say that doing so helps to protect against 'questionable research practices' (Isbell et al., 2022) and provides an audit trail by which we, as authors, can be held to account for our methods (Petticrew & Roberts, 2006; Gough, Oliver, & Thomas, 2012).

As a part of that process, we needed to choose an appropriate approach to appraising the quality of the eligible research. In Section 2.1.2 we remarked on the tendency for this important component of research synthesis to be omitted by systematic reviewers in Applied Linguistics, and how a systematic review without

this component is incomplete. It was important to us that we did not replicate this error in our own work. Moreover, we wanted to ensure that we did so in a way that was systematic, transparent and accountable.

For this, we turned to formalised appraisal tools to guide our decision making. The advantage of using published appraisal tools is that they provide a clear taxonomy, transparent decision-making guidance, and, therefore, an accountable approach to this component of the systematic review process. There are an enormous number of such tools available to reviewers, each of which tend to be suited to specific research designs. However, since we were not excluding studies based on their design (as long as they asserted causal relationships between using songs and substantive linguistic outcomes), we expected to find studies that use a variety of methods. Consequently, we chose to use the Mixed Methods Appraisal Tool (MMAT; Hong et al., 2018).

The MMAT allows reviewers to appraise studies that include multiple or 'mixed' data collection methods, such as combining quantitative measures of vocabulary acquisition and observational data of children's engagement with songs to draw conclusions about their effects. Its multifaceted framework facilitates the comparison of quality across studies that take different approaches to data collection. This was preferable to splitting papers into their respective qual/quant parts and ignoring their holistic nature when drawing conclusions, or to using multiple appraisal frameworks to assess quality and then trying to draw conclusions about the literature as a whole.

Despite its appropriateness to the task, using the MMAT does not come without challenges. As discussed in Section 2.1.4, the quality of reporting in primary studies is often a barrier to understanding exactly what authors did, and to extracting the data necessary to adequately synthesise a collection of related studies and confidently assess their quality. This raised an interesting ethical issue about whether a 'reporting quality' threshold should be considered as a criterion for inclusion. However, as our review aimed to provide a comprehensive overview of relevant research, we decided that reporting quality would not be a reason for exclusion. Understanding the methodological quality of included studies is important to understanding the strength of the gathered evidence, and the methodological state-of-the-art in the field generally.

Another challenge related to the provenance of the tool. As with much of the existing literature concerning systematic review methods, the MMAT was developed by healthcare researchers. We are of the opinion that robust design for asserting causal relationships is domain-general, not domain-specific. That is, the principles of robust causal design are independent of the specific interventions being evaluated. It does not matter whether you are comparing the relative effects of healthcare interventions or of educational interventions: aspects like baseline

equivalence of comparison groups, adequate scale, representativeness of the sample, and rates of attrition apply equally. However, when it came to submitting the review for publication, we found that not everyone shares this disposition. Peer reviewers were unhappy about the choice of a 'medical appraisal tool' to assess an educational body of research. In our view, they erroneously conflated 'medical' with 'basic science' (only a small part of healthcare research) rather than 'applied science' (a much larger aspect of healthcare research and methodologically much more similar to applied social science). These comments were a misunderstanding of how the MMAT can be applied (and indeed has been applied) to research from any paradigm or field. The principles of good research apply regardless of topic or context. A fair comparison is a fair comparison, and a transparent write-up is a transparent write-up. Nonetheless, it took months of persuasion to overcome this methodological clash of philosophies and to see the review published. This has highlighted for us the need for our field to produce trustworthiness appraisal tools designed by and for applied linguists in the hope that this will head off this type of *ad hominem* criticism and instead allow our field to concentrate on the quality of the research on the basis of widely understood and widely applicable methodological characteristics.

The final document was written up in full adherence to PRISMA 2020, and as such we hope that it addressed the concerns we have raised about reviews being only partially reported and questionably conducted. In addition, we shared our data, either in the published manuscript or as supplementary materials, including the decisions we made when appraising quality.

4. Alternative courses of action

Since conducting the review, we have been made aware of education-specific quality appraisal tools (Gorard, 2024; EEF, 2019). Should we need to update the review in the future, we may consider using one of these instead of the MMAT. This is not because we think the MMAT is inappropriate. It is because using a tool designed by and for education researchers might help to remove conceptual and philosophical barriers to communicating quality appraisal effectively, and thus allow for focus on the substantive message of the research rather than the (in our view, erroneous) criticism of the tools that helped produce that message, as described in Section 3. Alternatively, at the time, we might have considered creating a bespoke appraisal tool for the purposes of our systematic review, or using multiple tools, each for different designs (e.g. the Cochrane Risk of Bias2 tool (Cochrane, 2019) or ROBINS-I (Moore, Dwar, & Flemyng, 2023) for randomised trials and non-randomised comparisons, respectively). However, these courses of

action would have led to several issues that had potential to reduce the reliability of our review. Using multiple appraisal tools compromises the common frame of reference provided by the MMAT, and developing a bespoke tool would require rigorous validity and reliability testing, which was beyond the scope of CH's doctoral project.

Returning to the points made in Section 2.1.1, where we discussed the importance of user involvement in setting research questions, there is a good case for involving research users directly in the systematic review process. Gough, Oliver and Thomas (2012) devote a chapter of their book *An Introduction to Systematic Reviews* to demonstrating the importance of stakeholder perspectives and participation in reviews. In that chapter they outline ways in which stakeholders can (and should) be involved in the entire review process, from setting review questions, through defining meaningful outcomes, to disseminating findings. Again, the constraints of conducting a doctoral project precluded this level of stakeholder involvement at the time, but it is something we would hope to build into future reviews we undertake.

5. Conclusion

By undertaking to conduct an extensive and methodologically robust systematic review we were in a position to confidently address the uncertainty CH had witnessed surrounding strongly held practitioner beliefs that using songs and nursery rhymes to teach second or foreign languages will provide clear advantages for children's linguistic development, and the nature of the research that might substantiate this belief. While we did not find a clear answer to the question, we have produced reliable evidence about the state-of-the-art on this topic, and thus provide an evidence-based starting point for future related research. This bore fruit immediately, as it provided the basis for the intervention study CH conducted in the second phase of her project and thus allowed her to state with confidence that her research made a novel contribution (the *sine qua non* of doctoral research). We also learned that our contra-orthodox findings ran up against prevailing beliefs and biases, which necessitated an extensive period of discussion with reviewers to justify our findings. Primarily, the use of the MMAT raised questions about the suitability of a 'medical' appraisal tool for assessing research conducted in educational contexts. However, it is our belief that good research practice is independent of the subject matter, and that following a protocol to develop a review that can be replicated is a valuable part of the research process. We thus devised, in the published review and in collaboration with the reviewers and editors, a way of explaining the thinking behind our review process that will

potentially lead future systematic reviewers towards following clear protocols that reduce research waste, and thus increase the ethical and methodological impacts of systematic reviews in Applied Linguistics.

6. Questions for discussion

1. Do you agree that new research should be informed by the priorities of stakeholders, or is there a case for research for its own sake?
2. Are we correct to say that the characteristics of experimental design apply to all domains? Or is there something unique about experimental Applied Linguistics research that requires us to apply different standards?
3. Does Applied Linguistics need its own set of reporting guidelines for primary and secondary research, or can we borrow from other fields (e.g. CONSORT, PRISMA)?
4. Does Ioannidis' (2016) observation that his field is awash with "massive production of unnecessary, misleading, and conflicted systematic reviews" reflect the situation in Applied Linguistics?

Further reading

doi Chalmers, H., Brown, J. & Koryakina, A. (2024). Topics, publication patterns, and reporting quality in systematic reviews in language education. Lessons from the international database of education systematic reviews (IDESR). *Applied Linguistics Review*, *15*(4), 1645–1669.

This audit of systematic reviews in the field of language education analyses the contents of the International Database of Education Systematic Reviews (IDESR.org), for their adherence to PRISMA guidelines, and by proxy their methodological quality. It shows that adherence to PRISMA is extremely varied, and that the majority of systematic reviews fall well short of the highest standards of method and reporting. It offers suggestions for improving the state-of-the-art.

doi Isaacs, T., & Chalmers, H. (2025). Reducing 'avoidable research waste' in applied linguistics research: Lessons from healthcare research. *Language Teaching*, *58*(2), 214–231.

This paper synthesises thinking from healthcare, education, and Applied Linguistics on the concept of avoidable research waste. It defines five sources of research waste and describes ways in which researchers can maximise the likelihood that they avoid these. This includes promoting systematic reference to existing literature before conducting new primary and secondary research as a way to improve quality, relevance and impact.

doi Gorard, S. (2024). Judging the relative trustworthiness of research results: How to do it and why it matters. *Review of Education*, *12*(1), Article e3448.

This paper sets out the argument for why quality appraisal of primary research is important, indeed necessary, when conducting research syntheses that aim to inform policy and practice in the social sciences. It demonstrates why research synthesis is misleading when it does not take into account evidence quality and provides a framework for quality appraisal that can be used in Applied Linguistics.

AI use declaration statement

The authors confirm that no AI tools were used in the preparation, writing, formatting, and editing of this work.

References

Aloe, A. M., Dewidar, O., Hennessy, E. A., Pigott, T., Stewart, G., Welch, V., Wilson, D. B., Campbell MECCIR Working Group. (2024). Campbell standards: Modernizing Campbell's Methodologic Expectations for Campbell Collaboration Intervention Reviews (MECCIR). *Campbell Systematic Reviews*, *20*, e1445.

Aromataris, E., Lockwood, C., Porritt, K., Pilla, B., & Jordan, Z. (Eds.). (2024). *JBI manual for evidence synthesis*. JBI. https://synthesismanual.jbi.global.

Arslan, A. (2020). A systematic review on flipped learning in teaching English as a foreign or second language. *Journal of Language and Linguistic Studies*, *16*(2), 775–797.

Bruner, J. S. (1996). *The culture of education*. Harvard University Press.

Chalmers, H., Brown, J., & Koryakina, A. (2024). Topics, publication patterns, and reporting quality in systematic reviews in language education. Lessons from the international database of education systematic reviews (IDESR). *Applied Linguistics Review*, *15*(4), 1645–1669.

Chalmers, I., & Glasziou, P. (2009). Avoidable waste in the production and reporting of research evidence. *The Lancet*, *374*(9683), 86–89.

Clarke, M., Chalmers, I., Alderson, P., Hopewell, S. (2024). Reports of randomised control trials should begin and conclude with up-to-date systematic reviews of other relevant trials: A 25-year audit of the quality of trial reports. *Journal of the Royal Society of Medicine*, *117*(6), 212–216.

Cochrane (2019). *A revised tool to assess risk of bias in randomized trials (RoB 2)*. https://sites.google.com/site/riskofbiastool/welcome/rob-2-0-tool?authuser=0

Davis, G. M. (2017). Songs in the young learner classroom: A critical review of evidence. *ELT Journal*, *71*(4), 445–455.

Du, J., & Daniel, B. K. (2024). A systematic review of AI-powered chatbots in EFL speaking practice: Transforming language education. *Computers and Education: Artificial Intelligence*, *6*, 100230.

EEF (2019). *Classification of the security of findings from EEF evaluations*. https://d2tic4wvo1iusb.cloudfront.net/production/documents/evaluation/peer-review-process/Classifying_the_security_of_EEF_findings_2019.pdf?v=1745988135

EPPI Centre (2019a). *What is a systematic review?* https://eppi.ioe.ac.uk/cms/Default.aspx?tabid=67

EPPI Centre (2019b). *Methodological development*. https://eppi.ioe.ac.uk/cms/Default.aspx?tabid=1919#Approaches

Equator Network. (n.d.). *What is a reporting guideline?* https://www.equator-network.org/about-us/what-is-a-reporting-guideline/

Garton, S., Copland, F., & Burns, A. (2011). Investigating global practices in teaching English to young learners. *ELT Research Papers*, *11*. British Council

Gorard, S. (2015). *A proposal for judging the trustworthiness of research findings*. Durham Repository. https://durham-repository.worktribe.com/output/1415931

doi Gorard, S. (2024). Judging the relative trustworthiness of research results: How to do it and why it matters. *Review of Education*, *12*(1), Article e3448.

Gough, D., Oliver, S., & Thomas, J. (2012). *An introduction to systematic reviews*. Sage.

doi Hamilton, C., & Murphy, V.A. (2023). Folk pedagogy? Investigating how and why UK early years and primary teachers use songs with young learners. *Education 3–13*, *52*(8), 1488–1509.

doi Hamilton, C., Schulz, J., Chalmers, H., & Murphy, V.A. (2024). Investigating the substantive linguistic effects of using songs for teaching second or foreign languages to preschool, primary and secondary school learners: A systematic review of intervention research. *System*, *124*(103350).

doi Harris, J., & O'Leary, D. (2009). A third language at primary level in Ireland: An independent evaluation of the modern languages in primary schools initiative. In M. Nikolov (Ed.), *Early learning of modern foreign languages: Processes and outcomes* (pp. 1–14). Multilingual Matters.

Higgins, J.P.T., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M.J., & Welch, V.A. (Eds). (2024). *Cochrane handbook for systematic reviews of interventions* (Version 6.5, updated August 2024). Cochrane. https://www.training.cochrane.org/handbook

doi Hong, Q.N., Pluye, P., Fàbregues, S., et al. (2018). *Mixed methods appraisal tool (MMAT) Version 2018 User Guide*. McGill University.

doi Ioannidis, J.P.A. (2016). The mass production of redundant, misleading, and conflicting systematic reviews and meta-analyses. *The Milbank Quarterly* *94*(3), 485–514.

doi Isaacs, T., & Chalmers, H. (2025). Reducing 'avoidable research waste' in Applied Linguistics research: Lessons from healthcare research. *Language Teaching*, *58*(2), 214–231.

doi Isbell, D., Brown, D., Chen, M., Derick, D.J., Ghanem, R., Gutiérrez Arvizu, M.N., Schnur, E., Zhang, M., & Plonksy, L. (2022). Misconduct and questionable research practices: The ethics of quantitative data handling and reporting in Applied Linguistics. *The Modern Language Journal*, *106*(1), 172–195.

doi Macleod, M.R., Michie, S., Roberts, I., Dirnagl, U., Chalmers, I., Ioannidis, J.P., Al-Shahi Salman, R., Chan, A.W., & Glasziou, P. (2014). Biomedical research: Increasing value, reducing waste. *Lancet*, *383*(9912), 101–104.

Moher, D., Hopewell, S., Schulz, K., Montori, V., Gøtzsche, P., Devereaux, P. et al. (2010). CONSORT 2010 explanation and elaboration: Updated guidelines for reporting parallel group randomised trials. *International Journal of Surgery*, *10*(1), 28–55.

Moher, D., Shamseer, L., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., Shekelle, P., & Stewart, L.A. (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Systematic Reviews*, *4*(1), 1.

Moore, T.H., Dwan, K., & Flemyng, E. (2023). *ROBINS-I tool resources and reporting guidance*. Cochrane. https://methods.cochrane.org/sites/methods.cochrane.org/files/uploads/inline-files/Final_ROBINS-I%20author%20guidance_0.pdf

Nitami, D., & Santosa, M.H. (2024). A systematic literature review: The effectiveness of vocabulary size tests on students English vocabulary and writing skills. *LETS: Journal of Linguistics and English Teaching Studies*, *6*(1), 1–13.

Page, M.J., McKenzie, J.E., Bossuyt, P.M. et al. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372, n71.

Petticrew, M. & Roberts, H. (2006). *Systematic reviews in the social sciences: A practical guide*. Blackwell.

PRISMA (2025). *PRISMA Endorsers*. https://www.prisma-statement.org/prisma-endorsers

Sposet, B.A. (2008). *The role of music in second language acquisition; a bibliographic review of seventy years of research, 1937–2007*. Edwin Mellen Press.

Turnbull, H.W. (Ed.). (1959). *The correspondence of Isaac Newton: 1661–1675, Volume 1*. Published for the Royal Society at the University Press.

Vuogan, A., & Li, S. (2024). A systematic review of meta-analyses in second language research: Current practices, issues, and recommendations. *Applied Linguistics Review*, *15*(4), 1621–1644.

Werner, V. (2020). "Song-advantage" or "cost of singing"? A research synthesis of classroom-based intervention studies applying lyrics-based language teaching (1972–2019). *Journal of Second Language Teaching and Research*, *8*(1), 138–170.

CHAPTER 16

The opportunities and challenges of conducting a methodological review on the applications of Q methodology in Applied Linguistics

Nicola Morea & Hessameddin Ghanbar
Cambridge University Press & Assessment | Islamic Azad University

In this chapter, we reflect on the process of conducting a methodological review on the applications of Q methodology in Applied Linguistics. In particular, we focus on the methodological dilemmas and issues arising from such work, and how these have been addressed. For example, we consider issues around selecting qualifying empirical research, including the anticipated challenge of delimiting what counts as an application of this methodology, or distinguishing between an empirical study and a theoretical study that uses a case exemplar. We also reflect on the challenges of coding analytical approaches in a context characterised by frequent lack of methodological transparency in published research. The chapter concludes with a reflection on the lessons learnt during this process.

Keywords: Q methodology, Q-sort, methodological review, Applied Linguistics

Background

In this book chapter, we reflect on the experience of conducting a systematic review of the applications of Q methodology (Q) in Applied Linguistics (AL). The idea of mapping the use of Q in the field originated from the second author in July 2022; this resulted in a collaboration that lasted nearly 1.5 years and that led to the publication of our efforts in December 2023 on the journal *System* (Morea & Ghanbar, 2024).

This work can be considered a methodological review, and namely a systematic approach to "examine any methodological issues relating to the design, conduct and review of research studies and also evidence syntheses" (Munn et al.,

https://doi.org/10.1075/rmal.17.16mor
Available under the CC BY-NC-ND 4.0 license.© 2026 John Benjamins Publishing Company

2018, p. 5). The justification for this work stemmed from two trends in the field. On the one hand, we noticed a rise in empirical research in AL adopting Q to investigate subjectivity around language-related issues. In particular, we noticed a shift from initial theoretical and unsystematic reviews which introduced the methodology to the field (Irie, 2014; Li, 2022; Mahmoodi & Yousefi, 2021; Thumvichit, 2022a) to the emergence of an increasing number of empirical studies, particularly in areas related to language learners' and teachers' emotions (Fraschini & Park, 2022; Thumvichit, 2022b), motivation (e.g., Caruso & Fraschini, 2021; Fraschini & Caruso, 2019) and pedagogical beliefs (e.g., Lundberg, 2019a, 2019b). On the other hand, methodological reviews were also gaining momentum in the field (e.g., Aryadoust, Ng, & Sayama, 2021; Plonsky & Ghanbar, 2018); this was coupled with an increase in methodological reviews around Q methodology outside AL; such as in education (Lundberg et al., 2020), healthcare (Churruca et al., 2021), and across fields (Morea, 2022). Indeed, the time seemed ripe for such an exploration.

Q is a research methodology developed in the 20th century in the field of psychology with the aim of systematically studying human subjectivity (Stephenson, 1935, 1953). Operationally, it involves the design and administration of a sorting activity, during which a small number of research participants rank-order a large list of carefully selected statements on a pre-defined grid. These statements represent diverse perspectives on the topic under study. The sorting activity leads to the creation of a Q-sort, namely a grid populated with the Q-set items ranked in a way that best reflects each participant's viewpoints and experiences on the subject. The sample of completed Q-sorts is then analysed using a Q-specific analytical approach referred to as by-person factor analysis. This process results in the uncovering of a set of underlying factors reflecting shared configurations of viewpoints held by clusters of participants. Through a quantitative examination of each factor's structure (the factor array) — often triangulated with qualitative data from participant interviews conducted immediately after the sorting activity — a more in-depth description and interpretation of each factor is provided. Through this process, researchers systematically analyse and interpret the variety of participant viewpoints, assigning conceptual meaning and structure to emergent patterns. Through a combined analysis of Q-sort and interview data, they identify shared perspectives among participants aligned with each factor. For more information on Q, the authors recommend consulting Watts and Stenner's introductory book (2012) or reading the brief yet comprehensive primer by Brown (1993); an overview of each step of a Q study has also been provided in our recent methodological review on Q (Morea & Ghanbar, 2024) (see the *Further reading* section at the end of the chapter for more information).

Q methodology thus offers a distinct yet well-suited approach for investigating subjective phenomena — including beliefs, experiences, identity, motivation and emotions — all of which are central to AL research on language teaching, learning and acquisition. Being quite distinct from other quantitative and qualitative approaches for exploring subjectivity (e.g., questionnaires and psychometric scales, or interviews and focus groups), Q requires a range of decisions at all stages of a study, from the identification of research questions, participant recruitment, instrument design and refinement to data analysis and reporting. As a result, exploring and mapping the applications of Q in AL could not only shed light onto research domains in the field, but it would also allow us to evaluate its uses and, where necessary, provide recommendations for future applications of the methodology.

With this goal in mind, we embarked on a systematic database search to retrieve empirical Q research focusing on language-related matters, and to evaluate their contextual, methodological and data-analytical characteristics in order to understand how Q had so far served the field. In line with this, we did not specify any time period in our search, so that we could assess the potential growth of the methodology over time. Guided by the PRISMA guidelines (Page et al., 2021), we selected a set of qualifying criteria, defined the datasets to be consulted (as well as the search terms and strings), we drafted, piloted and refined a coding framework, and we finally conducted a database search, resulting in the identification of 55 eligible records.

The analysis of this sample of studies ($N = 55$) immediately brought to the light the rapid growth of Q in the field — as an example, more Q studies were published between 2019 and 2022 than in the previous twenty years. The review also showed that, whilst Q was being employed in a variety of areas within the field (e.g., language policy, language disorders, language education), most of the reviewed studies explored language learners' and teachers' cognition, emotions and motivation. When considering the methodological characteristics of the reviewed studies, a key finding was the conspicuous and frequent lack of information related to research design (e.g., any use of quality assessment procedures during Q-set development) and data-analytical decisions (e.g., factor extraction, rotation and retention procedures). Such finding clearly highlighted the need for more methodological transparency when publishing Q research in AL. We hypothesised that this trend may partially stem from limited familiarity with the methodology among journal editors and/or readerships, potentially leading to editorial decisions to restrict more technical methodological details. However, we argued that the relatively established status that Q methodology had gained in the field did not justify such approach any longer.

The above background information was intended to provide an overview of the rationale and key characteristics of our study. In the following sections, we examine the methodological challenges and dilemmas — both anticipated and unforeseen — encountered during the design and implementation of this review. Subsequently, we elucidate our approaches to these methodological dilemmas and critically examine potential alternative solutions. This discussion serves two goals: (1) to enhance readers' understanding of challenges inherent to methodological reviews, and (2) to propose actionable strategies for addressing such challenges.

Methodological and ethical dilemmas and issues

In this section, we present and discuss seven key dilemmas that emerged at different stages of our methodological synthesis on Q. Four of these were issues that we had anticipated and planned for during the design process, whilst the other three represent unanticipated dilemmas that we faced only in later phases of the review. Whilst each issue is presented below, the solutions we adopted to solve or reduce them will be the focus of discussion of the next section.

Anticipated issues

Defining AL research

In the early stages of designing our review, we anticipated several difficulties regarding the inclusion criteria and the retrieval of eligible records. In a review aimed at mapping the applications of Q methodology in AL, the first anticipated issue related to the challenge of defining what counted as Applied-Linguistic research (also refer to Oxley, 2026, for a similar dilemma in relation to 'English as an additional language'). A key objective of systematic reviews is to obtain a sample of studies approximating the population of relevant research around the target domain (Plonsky, Sudina, & Hu, 2021). The relevance of each potential study is typically assessed against a set of pre-defined inclusion or eligibility criteria. Whilst the need of specifying one inclusion criterion related to the field (as well as one related to the methodology being reviewed — more on this later) was immediately clear, delimiting such borders was not straightforward.

Whilst there seems to be agreement in characterising the field of AL as *broad, interdisciplinary* and concerned with real-world issues related to language, a consensus on its definition and scope has not been reached (Berns & Matsuda, 2006). With a lack of a shared definition and thus delimitation of the field, multiple potential approaches were possible. For example, one could embrace

a more restrictive approach and only consider specific macro-areas of AL, such as second/additional language research (thus excluding research into first language acquisition and education); alternatively, one could adopt a more all-encompassing approach that does not differentiate based on the nature of the language under scrutiny. Within the latter approach, one could draw a line between sub-fields and exclude areas of inquiry where language is but one of the foci (e.g., literacy acquisition and education, acculturation, language disorders), or rather consider any areas of inquiry where language represents at least one of the domains under scrutiny.

Defining Q research

A similar dilemma also emerged when trying to define what could be systematically considered as an application of Q methodology. Q being a complex methodology encompassing all stages of a study (i.e., research and instrument design, sampling, data collection and analysis, data interpretation), there was not a single way of operationalising what would qualify as a 'Q study'. For instance, it is not uncommon for the Q-sort method (also referred to as *Q-technique*) to be used within a study that would otherwise not align with the procedures of Q methodology (see, as an example, Wiefferink, Vermeij, & Uilenburg, 2012, excluded from our review). Moreover, some studies employing conventional Q methodological procedures for data collection and analysis may not fully align with — or even explicitly acknowledge — the epistemological foundations of the approach. Finally, how should non-standard applications of Q be treated, such as single-case and longitudinal research? We immediately recognised the challenge of establishing clear eligibility criteria to systematically distinguish between complete Q methodology studies and those merely influenced by its approach.

Longitudinal and experimental Q research

Another dilemma related to the above issue related to the treatment of longitudinal and experimental Q research, namely studies that employed Q as part of a within-subject design involving multiple phases of data administration and analyses over a period of time. The issue stemmed from how these studies would be coded in relation to each of our two research questions.

Our study had two aims; firstly, to map the applications of Q in the field of AL (e.g., tracking the rate of publications over the years, the publication type, the areas in the field where the methodology was used) and, secondly, to provide a rich description of the methodological and data-analytical characteristics of the retrieved studies (e.g., participant sample, item sample, number of factors extracted, criteria used for factor extraction, rotation and retention). Therefore, the complication of how to treat longitudinal Q research was due to the fact that

these studies involved multiple instrument administrations, datasets and analyses. As such, the unit of analysis differed between our two research questions. Specifically, a longitudinal study involving the administration of the same instrument (Q-set) at two time points to the same sample of persons (which may still differ due to participant withdrawal) would still count as a single unit for the purpose of the first research question, but as two units for the second question. Specifically, more than one sets of data could be coded for variables relevant to the second research question, such as sample size (which may or may not have differed at each time point) or number of factors extracted (which may or may not have differed at each factor analysis). In turn, this would result in an increased sample size when answering the second research question, as well as problems of dependence within this latter sample of studies.

Language-of-publication bias

A final anticipated dilemma concerned the language of publication of the eligible records. Whilst there is a tendency for systematic reviews to focus only on research published in English, the application of language restrictions during database search may undermine the systematicity of the review process and produce bias (Pieper & Puljak, 2021). Consequently, it was important to define a strategy in relation to the language of publication, and specifically to decide whether to adopt a more systematic approach (for example by systematically including only records in English or in specific languages) or to avoid applying any language restrictions during the database search (also see Aoyama, 2026).

Unanticipated issues

Three main unexpected issues emerged in subsequent phases of the project, namely during the screening, analysis and publication stages. As we did not foresee these dilemmas during the initial research design phase, each issue required an ad-hoc discussion within the research team until a consensus was reached.

Accessing full texts

During the screening phase — namely when assessing a set of retrieved studies against the inclusion criteria — one of the most frequent reasons for exclusion of a record was the impossibility of retrieving its full text (8/30, or 27%, of excluded records). This was due to paywalled publications for which we did not have institutional access, ranging from full books, book chapters, as well as journal articles. The impossibility of accessing a relatively large body of research risked introducing unwanted bias into our sample of eligible studies.

Unreported information

A second issue became apparent very early during the coding stage, and namely the frequent lack of reported information related to methodological and analytical decisions within our sample of eligible studies. Whilst we expected instances of unreported data, we were surprised by the frequency and importance of these omissions. As an example, in 34% ($n = 22$) of the 65 factor analyses conducted across our sample of 55 studies, no information was provided on the factor extraction method used (e.g., principal component analysis or factor analysis). Whilst this represents a finding in itself and did not necessarily require any remedy from the research team, it posed the question of the extent to which descriptive statistics from variables with a high rate of missing values may indeed approximate the true mean of the population of reference. Equally, it also led to discussions on whether and to what extent data not explicitly mentioned in the main body of the text could still be retrieved through other means, for example by analysing datasets and outputs provided in the form of appendices and supplementary materials.

Time gap between study retrieval and publication

A final unanticipated issue emerged during the article submission process and was related to the time gap between the study retrieval phase and the publication of our methodological review. By the time we received the reviewers' comments, approximately 15 months had passed. As a result (and in line with the near-exponential growth in published Q research in AL revealed in our study), new relevant research had been published in the meantime, which was inevitably not included in our review. In this context, should the original database search and resulting analysis be re-conducted and revised? If so, what about new, potentially eligible research that may appear during this revision, as well as peer-review and publication, processes? These were questions that we had to ask ourselves as we approached the publication of the review.

Course of action

In the previous section, we presented seven key issues that we encountered at different stages of the review, some which we had expected and planned for, and others that we had not initially foreseen. Here, we discuss how we addressed each of the seven issues. For clarity, these are reported in the same order as in the previous section. It is important to note that these courses of action were not the 'right' solutions to each problem, but rather one of many possible approaches that could have been adopted in the context of our project and in light of the research team's capacity.

Response to Issue One: Delimiting the field of AL

As previously mentioned, the lack of a shared definition of the field of AL required careful consideration of how we might operationalise what counted as AL research during the screening process. Our first step was to review and compare several definitions of AL from different sources, such as the "What is AILA" section on the International Association of Applied Linguistics website (AILA, n.d.), the "About Applied Linguistics" section on the American Association for Applied Linguistics website (AAAL, n.d.) and the "Aims & Scope" sections of the Research Methods in Applied Linguistics journal (RMAL, n.d.). As a result of this review, and in line with the aim of providing an overview of Q research that was as comprehensive as possible, we decided to embrace a broad definition of the field. Accordingly, the field-specific eligibility criterion was worded as: "the study must investigate a language-related phenomenon" (Morea & Ghanbar, 2024, p. 4). We thus included any empirical study where a language-related issue (regardless of the nature of the language, e.g., first or second) was the focus of investigation of at least one research question. This allowed us to include a wide range of studies concerned with language, such as research on early literacy instruction (Kearney-Edwards, 2016), on enhancing the quality of legal translations (Gao, 2019) and on voice-hallucinations among deaf people (Atkinson et al., 2007). On the other hand, we excluded any potentially eligible record where language was a related but secondary focus, such as research into international students' acculturation (e.g., Bang & Montgomery, 2013, not included in our review). Not all decisions were straightforward, and any records were flagged if they required a discussion within the research team to decide whether the inclusion criterion was met or not.

Response to Issue Two: Delimiting Q methodology

With regard to operationalising what could be considered as Q-methodological research, the availability of published methodological syntheses on Q meant that we could refer to how researchers had approached this issue. After a review of the eligibility criteria from published methodological reviews on Q research in various fields, we decided to follow the operationalisation adopted by Lundberg et al. (2020) and Churruca et al. (2021), by which a study was considered eligible if it employed *both* the Q-sort method for data collection and by-person factor analysis for data analysis. This approach is clearly pragmatic. It does not concern itself with assessing the extent to which a given study embraces the epistemological principles of the methodology, which, although important in principle, would have been difficult to implement in practice. Rather, it defines the two most important procedural characteristics of the methodology — the adoption of the

Q-sort method and the use of by-person factor analysis — as the minimum criteria for inclusion.

Response to Issue Three: Coding longitudinal and experimental Q research

Now that we had agreed on a clearly defined criterion for qualifying Q research, we could devote our attention to the treatment of non-conventional applications of Q methodology, and notably Q studies involving two or more phases of data collection and analyses. As established in our prior discussion, these studies would alter our unit of analysis. Consequently, we implemented distinct analytical approaches for these cases, with the specific treatment determined by the particular research question under examination. Regarding the first question (i.e., the mapping of the contextual characteristics of the reviewed studies), those records were simply treated as individual studies. Conversely, when coding and analysing the variables to answer our second research questions (i.e., the methodological and data-analytical characteristics of our sample of studies), we separated each record into multiple entries — or 'components' (Lundberg et al., 2020) — one for each factorial analysis conducted. In practice, this meant creating a separate row in our coding book for each additional analysis conducted within the same record, and only entering data for the variables relevant to our second research questions. This resulted in a sample size which was larger ($n = 65$) than the number of records selected for the reviews ($n = 55$).

Response to Issue Four: Language-of-publication bias

As previously mentioned, a possible source of bias in research syntheses is the application of language restrictions during the database search, resulting in the review of research only published in specified languages and, oftentimes, English. To try to tackle this issue prior to the database search, we decided not to impose any language restriction in our search string and to mobilise the multilingual repertoire of the two authors (spanning English, French, Italian and Persian), so as to include any study published in a language understood within the research team. However, we did not find potentially eligible studies (after screening of titles, key words and abstract) in languages known to the authors, and we had to exclude three records which appeared in languages unknown to us. At that point, we could only acknowledge this potential source of bias as a limitation of the study. Ultimately, we realise that our chosen approach did not produce different results from having included a language restriction either as an eligibility criterion or during the database search, so alternative approaches may be preferable in future reviews.

Response to Issue Five: Accessing potentially eligible records for full-text screening

A noticeable number of records eligible for full-text examination after screening of title, abstract and key words were behind a paywall ($n = 8$) and could not be accessed through our institutions. As the project was not supported by a research grant, it was not possible to purchase access to these publications. As a result, the only available approach was to exclude inaccessible records from further screening and acknowledge this in the review.

Response to Issue Six: Missing information in selected studies

The unanticipated issue of missing information around key methodological and data-analytical decisions in a noticeable proportion of our sample of studies did not, in practical terms, have any effects on our analytical procedures, as we coded these instances as missing data (i.e., coded as not reported). Nonetheless, to mitigate the risk of producing descriptive statistics that were biased due to the systematic exclusion of studies with missing information, we attempted to recover unreported data wherever possible. Specifically, when output files or data were made available (for example through repositories or as supplementary material), a consultation of these allowed us to confidently retrieve some of the otherwise unavailable information. One example of this was the retrieval of information on the treatment of confounded Q-sorts, an important decision in a Q study and one of the variables that we considered to answer our second research question. Confounded Q-sorts are Q-sorts with a significant factor loading on more than a factor, possibly resulting in higher inter-factor correlations when included in the analysis. Even if in a noticeable proportion of the analysed studies this information was not reported, we were able to determine the procedure that was used (e.g., whether confounded Q-sorts were systematically excluded or included, or whether an ad-hoc criterion was set to only exclude some of them) whenever output tables with factor loadings were made available. This resulted in a reduction of missing data for this variable to 37% of the total sample of studies. However, we only used this approach in those cases where the missing information could be retrieved with no degree of uncertainty. Whenever the available information did not provide a univocal answer (i.e., we could not confidently establish a specific methodological decision from the information we could retrieve), the data point was coded as not reported.

Response to Issue Seven: Keeping our review relevant

When our article was submitted to the journal *System*, approximately 12 months had passed since the study retrieval and inclusion phases. As a result, our paper had already aged even before being published. We were thus not surprised to see this issue being flagged by one of the reviewers, who recommended us to consider a way to update our review to make it more relevant by the time it would be published. Whilst, on the one hand, a time gap between database search and review publication is an inevitable feature of any research synthesis, we decided to act on the recommendation in recognition of the significance of such gap and the purpose of our review. We thus decided to conduct a second literature search using the same criteria of the original but adding a time restriction as to only include research published in the previous year. We were thus able to retrieve an additional 14 eligible studies. Yet, rather than revising our initial analysis and findings in light of this new sample, we decided to conduct a separate analysis of this second literature search. The results from this new review were discussed in a new sub-section of the discussion and compared with our original findings. This additional analysis therefore sought to examine whether the trends identified in our primary analysis (including publication rates and disciplinary distributions) persisted throughout the extended timeframe — a pattern we subsequently confirmed. Overall, whilst the decision to act on the reviewer's recommendation involved noticeable additional efforts, we considered it worth it as this addition not only increased the relevance of our study but also provided an opportunity to validate our findings.

Alternative courses of action

In the previous sections, we reflected upon some of the anticipated and unanticipated issues that we encountered during this project. We acknowledge that the way these issues were approached may not have been the only possible solutions, and many were motivated by the context in which the project took place. Now, we consider alternative courses of action and the potential impact they may have had for the research synthesis.

Considering the scope of our review, a different approach could have been adopted regarding the eligibility criteria for identifying Q studies. On the one hand, we could have employed a less strict approach for operationalising Q research, such as including any study employing the Q-sort method. On the other hand, we could have considered including non-empirical research, such as theoretical papers introducing the methodology to the field. These alternative

approaches would have resulted in a larger sample of eligible studies. Furthermore, another potential benefit would have been the possibility to track the frequency of different types of research over time. For example, one could have tracked different phases of Q research in AL, such as the emergence of empirical study from the foundations laid by initial theoretical works on the potential of Q for the field. Similarly, we could have investigated whether, and to what extent, applied linguists had become more or less conservative in their applications of the methodology over time (e.g., by embracing or moving away from traditional applications of the methodology in favour of methodological hybridisations). Nonetheless, these alternative approaches would not have been without limitations. Considering the lack of methodological reviews of Q in the field, as well as the considerable number of eligible records retrieved, a counterargument could have been made against the inclusion of studies that did not meet what Q methodologists may consider the minimum requirements to identify a Q-methodological study as such. By so doing, the review would have thus conflated Q methodological research with research *inspired by* Q methodology (see Wu & Morea, 2025, for an example of a Q-inspired research design). Furthermore, these studies could not have contributed to answering the second research question ("what are the Q-specific characteristics of the selected studies?"), due to the lack of empirical data and/or Q-specific data-analytical procedures. Ultimately, whilst acknowledging that other researchers might have favoured this alternative approach, we consider our chosen course of action valid in light of the study's research objectives and the specific context in which the review was conducted.

The second alternative scenario we wish to consider is related to the availability of funding to support the review project. Our project was not funded through a research grant or other forms of funding, which posed some limitations on our ability to access potentially eligible studies. As previously mentioned, funding availability would have allowed us to purchase paywalled records for which we did not have institutional access. Additionally, we could have adopted a different strategy for accessing records in languages not known to the research team. One strategy could have been to employ research assistants proficient in the target languages or to utilise professional translation services. However, it should be acknowledged that, as the search terms and string were in English, this alternative approach might have only reduced, rather than removed, the bias from an overrepresentation of research published in English.

Finally, since the publication of our methodological review, field-specific frameworks for conducting rigorous research syntheses have become available. In particular, the Synthesis Methods and Reporting Tool (SMART) (Chong, 2025), a framework specifically designed to guide research syntheses in the field of AL, could have been adopted to guide our project across all its stages. Although a pre-

print of our publication was made freely available on the SSRN platform, an open-access, peer-reviewed protocol could have been submitted prior to the database search, detailing the type and aim of the project, its research questions and timeline, and the intended approach for database search and reporting. Even though we found PRISMA (Page et al., 2021) a useful protocol for informing our search strategy and reporting, the SMART framework would have ensured alignment with other field-specific research syntheses.

Conclusion

In this chapter, we reflected on the anticipated and unanticipated challenges encountered when conducting a methodological synthesis Q methodology in AL. These included defining inclusion criteria that effectively delimited both the target field and methodology under scrutiny. We also reflected on sources of bias in retrieving a representative sample of studies aligned with the inclusion criteria due to difficulties in accessing some records and in reviewing research published in languages other than English.

Furthermore, methodological reviews pose challenges that distinguish them from other types of secondary research. For example, although research syntheses involve the creation of research teams able to cover multiple areas of expertise (e.g., on the target domain of the review and on the research synthesis methodology itself, see Chong, 2025), methodological synthesis additionally require specialism in the methodology to be reviewed. In this regard, we believe that the availability of field-specific frameworks for conducting different types of research syntheses such as SMART will greatly help researchers in conducting this relatively novel form of research synthesis.

We hope that this chapter has served the two purposes which it purported to achieve: first, by offering valuable insights into the methodological review process, including strategies for addressing both anticipated and emergent challenges; second, by potentially stimulating greater interest among applied linguists in utilising Q methodology as an innovative approach for examining language-related phenomena.

Questions for discussion

In consideration of the anticipated and unanticipated dilemmas that we encountered at different stages of our review, we propose five questions that researchers interested in conducting a methodological review in AL should consider:

1. How would you delimit/operationalise the target domain/field (e.g., Applied Linguistics)?
2. How would you delimit/operationalise the target research method and/or methodology?
3. How could you limit any language-related selection bias?
4. How would you approach methodological information recurrently absent in your sample of selected studies?
5. What actions (if any) would you take in case of a significant time gap between the database search and the review's publication?

Further reading

Below, we propose what we consider to be four essential references for exploring Q methodology, both within and beyond the field of AL.

Watts, S., & Stenner, P. (2012). *Doing Q methodology research: Theory, method and interpretation*. Sage.

This book is a classic introduction to Q methodology. It is an accessible and simple read that nonetheless provides an extensive coverage of both the epistemological and methodological principles behind Q and the procedural characteristics of a typical Q study.

Brown, S. R. (1993). A primer on Q methodology. *Operant Subjectivity*, *16*, 91–138.

This freely accessible article effectively exemplifies the data-analytical stages of a Q study through a hypothetical example. Particularly useful to understand the process of Q-sort correlation and by-person factor analysis.

Lundberg, A. (2019a). Teachers' beliefs about multilingualism: Findings from Q method research. *Current Issues in Language Planning*, *20*(3), 266–283.

In our opinion, this empirical study exemplifies the potential of Q methodology for AL research. Q methodology has been employed here to explore the beliefs about multilingualism of a sample of primary school teachers in Sweden, revealing consensus as well as tensions among teachers' views of multilingualism and their professional beliefs about pedagogical approaches around linguistic inclusion.

Morea, N. (2022). Investigating change in subjectivity: The analysis of Q-sorts in longitudinal research. *Research Methods in Applied Linguistics*, *1*(3),

This methodological review investigates the applications of Q in longitudinal and (quasi-)experimental research across fields. It provides a practical overview of the key stages of a traditional Q study before reviewing the characteristics of studies that have utilised the methodology as part of a within-subject research design. The study also proposes and exemplifies a multi-method approach for the analysis of Q-sorts collected from the same participants at multiple time points.

AI use declaration statement

The authors confirm that no AI tools were used in the preparation, writing, formatting, and editing of this work.

References

American Association for Applied Linguistics (n.d.). *About Applied Linguistics*. American Association for Applied Linguistics. Retrieved 10 March 2025 from https://www.aaal.org/about-applied-linguistics

Aoyama, T. (2026). Challenges and practical solutions in applied linguistics systematic review: A reflective perspective. In S.Q. Chong (Ed.), *Research synthesis methodologies in applied linguistics*. John Benjamins. (this volume)

doi Aryadoust, V., Ng, L.Y., & Sayama, H. (2021). A comprehensive review of Rasch measurement in language assessment: Recommendations and guidelines for research. *Language Testing*, *38*(1), 6–40.

Association Internationale de Linguistique Appliquée (n.d.). *What is AILA*. International Association of Applied Linguistics. Retrieved 20 March 2025 from https://aila.info

doi Atkinson, J.R., Gleeson, K., Cromwell, J., & O'Rourke, S. (2007). Exploring the perceptual characteristics of voice-hallucinations in deaf people. *Cognitive Neuropsychiatry*, *12*(4), 339–361.

doi Bang, H., & Montgomery, D. (2013). Understanding international graduate students' acculturation using Q methodology. *Journal of College Student Development* *54*(4), 343–360.

doi Berns, M., & Matsuda, K. (2006). Applied linguistics: Overview and history. In K. Brown (Ed.), *Encyclopedia of language & linguistics* (2nd ed., pp. 394–405), Elsevier.

doi Brown, S.R. (1993). A primer on Q methodology. *Operant Subjectivity*, *16*, 91–138.

doi Caruso, M., & Fraschini, N. (2021). A Q methodology study into vision of Italian L2 university students: An Australian perspective. *The Modern Language Journal*, *105*(2), 552–568.

doi Chong, S.W. (2025). Synthesis Methods and Reporting Tool (SMART) for research syntheses in Applied Linguistics. *Research Synthesis in Applied Linguistics*, *1*(1), 17–38.

doi Churruca, K., Ludlow, K., Wu, W., Gibbons, K., Nguyen, H.M., Ellis, L.A. et al. (2021). A scoping review of Q-methodology in healthcare research. *BMC Medical Research Methodology*, *21*(1).

doi Fraschini, N., & Caruso, M. (2019). "I can see myself…" A Q methodology study on self vision of Korean language learners. *System*, *87*, 1–12.

doi Fraschini, N., & Park, H. (2022). A Q methodology study to explore Korean as a second language undergraduate student-teachers' anxiety. *International Journal of Educational Research Open*, *3*.

Gao, L. (2019). A two-dimensional hypothetical model for legal translation competence enhancement. *Current Trends in Translation Teaching and Learning E*, *6*, 269–305.

doi Irie, K. (2014). Q methodology for post-social-turn research in SLA. *Studies in Second Language Learning and Teaching*, *4*(1), 13–32.

Kearney-Edwards, K. B. (2016). Beliefs of urban kindergarten teachers: Connecting early literacy instruction to families' literacy practices (Doctoral dissertation). George Mason University. ProQuest Dissertations & Theses Global.

Li, X. (2022). The significance of Q-methodology as an innovative method for the investigation of affective variables in second language acquisition. *Frontiers in Psychology*, *13*.

Lundberg, A. (2019a). Teachers' beliefs about multilingualism: Findings from Q method research. *Current Issues in Language Planning*, *20*(3), 266–283.

Lundberg, A. (2019b). Teachers' viewpoints about an educational reform concerning multilingualism in German-speaking Switzerland. *Learning and Instruction*, *64*.

Lundberg, A., de Leeuw, R., & Aliani, R. (2020). Using Q methodology: Sorting out subjectivity in educational research. *Educational Research Review*, *31*.

Mahmoodi, M. H., & Yousefi, M. (2021). Second language motivation research 2010–2019: A synthetic exploration. *The Language Learning Journal*, *50*(3), 273–296.

Morea, N. (2022). Investigating change in subjectivity: The analysis of Q-sorts in longitudinal research. *Research Methods in Applied Linguistics*, *1*(3).

Morea, N., & Ghanbar, H. (2024). Q methodology in applied linguistics: A systematic research synthesis. *System*, *120*.

Munn, Z., Stern, C., Aromataris, E., Lockwood, C., & Jordan, Z. (2018). What kind of systematic review should I conduct? A proposed typology and guidance for systematic reviewers in the medical and health sciences. *BMC Medical Research Methodology*, *18*(5).

Oxley, E. (2026). Conducting systematic reviews in applied linguistics: A case study and critical reflection. In S. Q. Chong (Ed.), *Research synthesis methodologies in Applied Linguistics*. John Benjamins. (this volume)

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, C. D., et al. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, p. 372.

Pieper, D., & Puljak, L. (2021). Language restrictions in systematic reviews should not be imposed in the search strategy but in the eligibility criteria if necessary. *Journal of Clinical Epidemiology*, *132*. 146–147.

Plonsky, L., & Ghanbar, H. (2018). Multiple regression in L2 research: A methodological synthesis and guide to interpreting R2 values. *The Modern Language Journal*, *102*(4), 713–731.

Plonsky, L., Sudina, E., & Hu, Y. (2021). Applying meta-analysis to research on bilingualism: An introduction. *Bilingualism: Language and Cognition*, *24*(5), 819–824.

Research Methods in Applied Linguistics (n. d.). *Aims and scope*. Retrieved 20 March 2025 from https://www.sciencedirect.com/journal/research-methods-in-applied-linguistics/about/aims-and-scope

Stephenson, W. (1953). *The study of behaviour; Q-technique and its methodology*. The University of Chicago Press.

Stephenson, W. (1935). Technique of factor analysis. *Nature*, *136*, p. 297.

Thumvichit, A. (2022a). Illuminating insights into subjectivity: Q as a methodology in applied linguistics research. *Applied Linguistics Review*, 15(2), 531–554.

Thumvichit, A. (2022b). Enjoyment in language teaching: A study into EFL teachers' subjectivities. *International Review of Applied Linguistics in Language Teaching*, *62*(2), 623–649.

Watts, S., & Stenner, P. (2012). *Doing Q methodology research: Theory, method and interpretation*. Sage.

Wiefferink, C. H., Vermeij, B. A. M., & Uilenburg, N. (2012). Family counseling in the Netherlands for Turkish-origin parents of deaf children with a cochlear implant. *American Annals of the Deaf*, *156*(5), 459–468.

Wu, X., & Morea, N. (2025). Putting subjectivity in time: Qualitatively examining multilingual identity development via a Q-inspired longitudinal design. *Research Methods in Applied Linguistics*, *4*(1).

SECTION 8

Emergent and interdisciplinary practices in research synthesis

CHAPTER 17

Human reviewers in an AI era

A methodological reflection on research synthesis

Dongxia Nie[1], Qi Liu[2], Dominique Yixue Jiang[3], Rui Yan[4], Yi Zhang[5], Xin Fang[6], Jingying Hu[7], Xuechun Huang[8] & Jing Yun[9]

[1] University of Oxford | [2] University of Leicester | [3] University College London | [4] University of Bath | [5] University of Warwick | [6] City University of Hong Kong | [7] Purdue University | [8] University of Leeds | [9] University of St Andrews

This chapter presents a methodological reflection on a large-scale meta-review commissioned by the British Association for Applied Linguistics. Drawing on reflections from an early-career researcher perspective, we discussed the interplay between human reviewers and AI-supported tools in research synthesis in Applied Linguistics. The chapter is organised around four themes: methodological robustness, workflow design, collaboration, and researcher development, with a particular focus on the role of human reviewers and AI-assisted tools and platforms in supporting these aspects. Specifically, we reflected on how automatic tools enhanced efficiency of our work and methodological rigour of the meta-review, while human judgement remained central to meaning-making and decision-making.

Keywords: methodological reflection, meta-review, AI tools, research synthesis, Applied Linguistics

1. Background

In September 2024, the British Association for Applied Linguistics (BAAL) commissioned a large-scale meta-review to address two fundamental questions: What is Applied Linguistics? and What does Applied Linguistics do? A research team was therefore formed to undertake this task. The team is led by Principal Investigator Professor Sin-Wang Chong, with Co-Investigators Dongxia Nie and Qi Liu, and supported by a College of Reviewers and an international advisory board representing diverse expertise in the field of Applied Linguistics. The primary

https://doi.org/10.1075/rmal.17.17nie
Available under the CC BY-NC-ND 4.0 license.© 2026 John Benjamins Publishing Company

objective of the project is to map the intellectual landscape and methodological practices of Applied Linguistics that examine how languages have been taught, learned, used, and researched.

Given the scale, diversity, and interdisciplinary nature of Applied Linguistics, a comprehensive analysis at the level of primary studies would be prohibitively resource-intensive and methodologically complex, with a risk of partial or fragmented representation. We therefore opt for a meta-review approach, which is also referred to as an "umbrella review", "review of reviews" or "meta systematic review" that focuses on synthesising existing research syntheses rather than primary studies (Chong, 2025; Chong & Plonsky, 2024). This approach is feasible because the number of research syntheses is more manageable than the volume of primary studies, with each review representing a body of primary research. As such, it enables a higher-level analysis of key areas and thematic patterns within the field. It is also appropriate in that Applied Linguistics has matured to a stage where a sufficient body of research syntheses has been produced to support a substantive and systematic meta-review.

To operationalise these aims, we designed a comprehensive search and screening procedure to identify relevant research syntheses in Applied Linguistics (Chong et al., 2025). The final data search was conducted on 5 March 2025 across four databases: Scopus, Web of Science, EBSCO, and APA PsycINFO. As a result, 10,858 records were identified and exported from the databases. Rayyan (Ouzzani et al., 2016) was then used to help remove 3,581 duplicates, leaving 7,277 records for title and abstract screening. At this stage, 5,014 records were excluded because they did not meet at least one of the inclusion criteria. Full-text screening was conducted via EPPI Reviewer (Thomas et al., 2023). We ended up excluding another 527 records after closer examination. Finally, 1,736 articles were retained for data extraction and synthesis. At the time of writing, the first round of data extraction is now nearing completion. Each article has been coded for its review type, broad Applied Linguistics category, and specific Applied Linguistics topic. The next step will be to analyse the data and present the results for the first stage.

The following sections are collective reflections from the team regarding various methodological stages up to data extraction. In addition to the procedural workflow, the project generated methodological insights that extend beyond this specific meta-review. A key insight was the distinctive advantage of human–AI collaboration in large-scale research synthesis. In recent years, the capabilities of AI-driven tools have expanded rapidly. Tools such as Elicit (Elicit, 2025) demonstrate that automation now extends beyond basic text processing. Instead of merely retrieving documents, these systems can analyse titles and abstracts, cluster studies by similarity, generate preliminary relevance predictions, and identify information from the texts based on predefined prompts. As AI tools become increasingly

capable of supporting multiple stages of research synthesis (Abogunrin et al., 2025; Marshall & Wallace, 2019), some have questioned whether human reviewers will remain central to the process (Scherbakov et al., 2025).

Our experience suggests otherwise. Across all stages of the meta-review, we made extensive use of AI-assisted tools, such as Rayyan for title and abstract screening and EPPI Reviewer for full-text screening and coding. We also plan to use EPPI Visualiser (Digital Solution Foundry and EPPI Centre, 2023) to present our results and build a living database. They have undoubtedly streamlined many aspects of this meta-review, including deduplication, relevance suggestions, and automated agreement checking. At the individual level, team members also used tools like ChatGPT and Elicit to help identify and crosscheck information. However, the team agreed that human involvement remained not only unavoidable but indispensable. Therefore, the question is not whether human reviewers could be replaced, but why they remain essential even as AI capabilities expand. The sections that follow discuss four interrelated themes that emerged from our methodological reflections: the role of human judgement in ensuring robustness, the structuring of workflows, the value of collaboration, and the development of human reviewers. As the team was composed primarily of early career researchers working under senior guidance, the reflections are grounded in an early-career researcher perspective.

2. Reflections from the team

2.1 Reflection on methodological robustness: Why human judgement mattered

Research synthesis is built upon principles of comprehensiveness, systematicity, and transparency (Plonsky et al., 2025), and in our project, a range of automatic tools played an important role in supporting these aims. In terms of comprehensiveness, by enabling large-scale data retrieval, detecting near-duplicate records, and highlighting potentially relevant information in the screening process, these tools helped reduce the risk of missing or prematurely excluding relevant articles. With respect to systematicity, the use of structured platforms such as Rayyan and EPPI Reviewer enforced consistent workflows across all reviewers, standardised the application of screening criteria, and ensured that records were handled in a uniform and reproducible manner. For transparency, the aforementioned platforms automatically documented every inclusion and exclusion decision, categorised reasons for exclusion, and generated agreement reports among reviewers that made areas of interpretive divergence visible to the entire team. Thanks to

these tools, our screening process remained fully traceable and independently verifiable.

Taken together, these automatic tools, especially their AI-supported features, enhanced the methodological robustness of the review. However, under the specific context of Applied Linguistics, an interdisciplinary field that values socially situated, reflexive, and context-sensitive interpretations of language, AI alone could not ensure conceptual clarity or relevance, as generative systems tend to reproduce normative language patterns embedded in their training data (Révész et al., 2025). Many records in our review required interpretive decisions that involved assessing how language was positioned within a study: whether as a central construct, a methodological tool, or an incidental variable embedded within clinical, cognitive, or technological research. In some cases, determining the fit of a study for inclusion required not only procedural consistency, but also nuanced reading and theoretical understanding. Here, human interpretation was indispensable, as reviewers with diverse expertise can make contextually informed decisions that no automated filter could reliably replicate.

Human expertise also played an essential role in evaluating the outputs produced by AI tools (Qureshi et al., 2023). While automated suggestions were helpful for flagging possible exclusions or identifying similarities, reviewers still needed to check these outputs for accuracy, relevance, and alignment with the conceptual aims of the meta-review. AI could accelerate processes and reduce mechanical workload, but it could not determine what qualifies as an Applied Linguistics research synthesis, nor could it effectively interpret epistemological orientation or disciplinary contribution (Alaqlobi et al., 2024; see also Chapter 13, this volume). These decisions depended on shared professional knowledge and the interpretive calibration developed through team discussions and multiple rounds of trial screening (see also Chapters 5 and 6, this volume, for discussions on methodological rigour and mapping in bibliometric reviews).

In this project, therefore, human and machine contributions were complementary. AI-supported tools helped us manage an exceptionally large dataset and avoid common human errors, but the actual decisions and the reasoning behind them depended on disciplinary knowledge, contextual interpretation, and collective discussion of human reviewers. The tools did not determine whether a study should be included or which category it should be coded into; they mainly supported reviewers in applying the criteria more consistently and efficiently.

The reflections below illustrate how team members experienced this interplay in practice. Notably, reviewers were not prompted to comment on this specific theme; rather, it emerged organically across the team's accounts as a shared and recurrent insight.

I also learned the usefulness of digital platforms. It was the first time that I tried Rayyan and EPPI Reviewer, and they turned out to be very helpful for screening. Compared with Excel that I used to employ for screening, features on these two platforms, such as writing reasons for inclusion and exclusion and comparing decisions across team members, improved efficiency, transparency and accuracy. Now I see their value and may consider using such specialised tools in my future synthesis. — Rui Yan

I struggled to make decisions during the screening stage as many studies were interdisciplinary. Particularly, some articles included language but only as a secondary focus. For example, I was unsure whether a study investigating cognitive development in children with language impairments should be included. It was relevant to language, but not the main focus of the article. This was one of the main challenges I faced at the screening stage. I brought this question to a meeting with Sin-Wang and Dongxia. We agreed that there is no right or wrong answer, only different interpretations, so we need to reach a consensus as a team. To achieve this, we went through six rounds of trials, during which many other issues arose, and we went back and forth on how to draw the line for defining Applied Linguistics. — Qi Liu

During the title and abstract screening stage, I initially relied on identifying key conceptual markers — such as terminology related to linguistic constructs, population descriptors, and research contexts — to exclude articles clearly outside the scope, particularly those from computational linguistics or clinical medicine. However, more nuanced challenges emerged with studies situated at the interface of language and therapeutic practice. These often involved language-related phenomena (e.g., processing or discourse patterns) but were framed in service of clinical, rather than linguistic, goals. Over time, I refined my strategy to move beyond keyword recognition and attend more carefully to research aims and epistemological orientation. This shift — towards assessing a study's underlying contribution to applied linguistic inquiry — proved crucial in making informed inclusion decisions. — Xin Fang

I vividly remember that during our first trial of title and abstract screening, the agreement rate was below 50%, even though the team used the same inclusion/exclusion criteria. This was because criteria such as "language as the main variable" involved subjective judgements about what constituted "main", leading to significant variation in our decisions. Through detailed discussions of concrete borderline cases and additional rounds of trials, we gradually unified the interpretation of key criteria. This collective calibration process aligned our decisions and significantly increased the agreement rate. Through this experience, I learned that methodological robustness is achieved not merely by having standardised criteria, but by developing shared understandings through human judgement.
— Dominique Yixue Jiang

> One of the most striking aspects of this project was the sheer scale of the screening task. With thousands of records pulled from multiple databases (Scopus, Web of Science, APA PsycINFO, EBSCO), it quickly became clear that defining scope was not straightforward. Applied Linguistics is an expansive field, and deciding whether certain reviews, such as those focusing on discourse analysis in medical contexts or NLP applications, should count as "Applied Linguistics" was often far from obvious. — Jingying Hu

> As we would expect, Applied Linguistics is a broad field. It seemed easy to decide if a study is within our field. However, we often found that the research was not primarily intended for applied linguists, but for clinical practitioners. On the one hand, our team realised that language was a crucial component of many studies. On the other hand, we needed to decide in which studies language was not the focus. Throughout the process, we deepened our understanding of the field by negotiating the boundaries of Applied Linguistics, which were precisely shaped by the research we reviewed. — Xuechun Huang

In the reflections, reviewers described disagreement, uncertainty, revision, recalibration, and gradual convergence. In a large-scale meta-review in an interdisciplinary field, we see these features not as weaknesses of the method, but as evidence of robustness in the making. In this process, AI tools helped enhance efficiency and consistency, but meaning-making was essentially realised by human reviewers.

2.2 Reflection on workflow: Automatic tools as a support, not a substitute

A second theme that emerged from our reflection relates to how the principles outlined above were achieved through the design and coordination of our workflow. This theme concerns the organisational infrastructure that enabled the review to be both scalable and rigorous. From the outset, we aimed to build a process that was both structured and adaptable, one that makes good use of automatic tools but still relies on human decision, communication and coordination. The workflow began with a comprehensive database search, after which all retrieved records were imported into a shared Rayyan project. Reviewers were added to the project and given access to the dataset, enabling a coordinated screening environment.

A key early task was to establish robust and operationalisable screening criteria. The investigators conducted six iterative rounds of pilot screening between December 2024 and January 2025 to refine the Inclusion/Exclusion Checklist. Two additional rounds of whole-team trial screening were then conducted to establish the final calibration of decision rules for both title/abstract and full-text stages. These iterative exercises were essential for identifying ambiguous cases, refining definitions, and aligning reviewers' interpretive approaches before formal

screening began. After that, we finalised the protocol and translated the criteria into practical, directly applicable rules within Rayyan. Each exclusion reason was standardised so reviewers could apply them consistently. In parallel, we developed a step-by-step checklist (see Appendix A) that outlined the reasoning process reviewers could follow when examining a study: what to look at first, what to verify next, what exceptions to consider, and how to move forward depending on each decision. The checklist served as a cognitive scaffold that helped reviewers use the same interpretive sequence even when dealing with ambiguous or complex cases.

Moving forward to full-text screening, we migrated to EPPI Reviewer, a platform that several team members found particularly efficient for documenting decisions. All studies marked as "include" at the title/abstract screening stage were exported from Rayyan in a single batch and then imported into EPPI Reviewer, together with the tags and notes assigned during the earlier stage. EPPI Reviewer's interface allowed us to record exclusion reasons with greater precision, aided by features that highlight relevant segments of the article as reviewers make decisions. It not only strengthened transparency in the full-text screening process but also laid the essential groundwork for the subsequent data extraction stage.

Before data extraction began, we developed a detailed coding scheme to break down our research questions into concrete units of information. The purpose of this scheme was to translate abstract, conceptual aspects of the meta-review into specific items that could be consistently extracted from the included studies. This required us to consider how different types of reviews report their research aims, contexts, methods, and findings, and what forms these elements might take in diverse and sometimes unconventional reporting formats. In other words, we needed to anticipate the multiple ways in which the same piece of information could be expressed and convert these possibilities into operationalisable coding categories.

While automatic platforms provided structured grounding to execute the workflow, the implementation relied fundamentally on human expertise. A major challenge was ensuring that decisions were consistent among reviewers, while still allowing the flexibility needed to handle the diversity of research syntheses included in the meta-review. Authors vary widely in how they report their methods and results. Some papers provided minimal information, others lacked abstracts altogether, and many presented unusual or hybrid formats. These nuances cannot be captured through a single automated process or rigid rule set, nor can they be resolved by platform functionalities alone.

Developing a protocol at the outset, refining the inclusion/exclusion criteria before screening, creating a practical checklist to guide screening decisions, and building a comprehensive coding scheme for data extraction all served to manage

this inherent variability by adding structured layers to the workflow. In addition, training and calibration exercises were used to align our interpretations and establish shared expectations at every stage. We hope to bring as much regularity as possible to a dataset characterised by irregularities (see also Chapter 7, this volume, for technical resolutions to methodological dilemmas in meta-analysis). Yet we also recognised that not all irregularities can or should be eliminated in research synthesis (Hampson & McKinley, 2024). Certain decisions required contextual reading, methodological sensitivity, and an understanding of how different review genres operate. In this respect, human involvement was indispensable.

Below are reflections from the team that shed light on how this workflow unfolded in practice and how reviewers navigated uncertainties in the process and gradually developed confidence in their work.

> At the beginning, the screening process felt messy for me because there were many criteria to consider (e.g., research synthesis, topic). After the six rounds of trials, I established my own screening procedure, such as first looking at whether the article was a research synthesis, then examining the role of language, and so on. Once a consensus was reached in the initial team, we introduced the project to other team members. After several rounds of Q&A, a checklist for screening was proposed by Dongxia. It resembled a computer processing flow, which provided a clear and systematic way of thinking for screening. I would adopt this format in my future projects, as following the same thinking process improves consistency. The checklist also highlighted the key checking points, which saved us time in getting familiar with the checklist. When everything is emphasised, nothing stands out. — Qi Liu

> Based on the shared understanding, a step-by-step checklist streamlining the screening process was developed and shared among team members, which addressed the challenge posed by a large quantity of entries. The checklist simulated the decision-making process and structured criteria into steps. It significantly increased my work efficiency as it broke the process into actionable steps, and ensured my consistency in applying the criteria throughout the process. Practices of creating and sharing easy-to-follow protocols or guidelines like the checklist have been implemented throughout the process. When built on team consensus and followed by all team members, they also contributed to reliability within the team. — Yi Zhang

> At the beginning of the abstract screening stage, I found the process overwhelming. During training, we were told that a few minutes should be enough to screen an article; however, in reality, I sometimes spent nearly an hour deliberating over borderline cases. Our first blind trial also resulted in a very low agreement rate across the team, which was particularly concerning given that we needed to screen more than 7,000 entries in total. However, thanks to repeated practice,

joint calibration sessions, and above all, the checklist developed by Dongxia, we gradually established shared clarity about the inclusion/exclusion criteria. A key principle that emerged at this stage was "include if in doubt", which ensured that potentially relevant studies were not prematurely excluded and gave us confidence to handle borderline cases more decisively. By the later stages, I could make decisions in well under five minutes, and what had initially felt overwhelming eventually became manageable. Upon completion of abstract screening, I entered the full-text screening stage with greater confidence, and the process felt much smoother, which perhaps best reflects how much the workflow had improved.

— Dominique Yixue Jiang

Doing two rounds of screening for this meta-review turned out to be a rich learning experience. At first, I thought screening would be straightforward, just going through titles, abstracts, as well as full-texts, and deciding whether they fit or not. But once I began to work, I realised how much planning, discussion, and teamwork is needed to do it properly and consistently. One of the first things I learned was the importance of pre-screening online training. Before we officially started screening, we tried out the screening platforms with a small set of studies, and that helped us get familiar with the necessary technical steps. For example, I figured out how to tag or take notes on Rayyan. It seemed to be a trivial issue, but without the training, I would have run into confusion later on when dealing with a larger number of studies. This made me always keep in mind that a trial or a pilot is important before doing research synthesis, since it equips team members with necessary skills and knowledge and therefore saves time in the long run. Another big takeaway was the usefulness of having a clear protocol, which streamlined the screening process and significantly improved its efficiency. When I encountered articles that were difficult to make decisions, such as whether to include articles from medical fields that still had relevance for teaching practitioners, the protocol gave us a reference point. Thanks to this detailed and updated protocol, we reached a consensus rather than making inconsistent choices.

— Rui Yan

Although all team members identified with the field, our individual research backgrounds and sub-disciplinary focus varied substantially. Initially, I mistakenly excluded some articles that, while unfamiliar to me, fell well within the broader applied linguistic tradition. This prompted an ongoing process of reconceptualizing the field — not just cognitively, but practically, within the screening process itself.

— Xin Fang

Our team can offer two good practices to the wider community of researchers. The first is standardising the inclusion and exclusion process in Rayyan and EPPI Reviewer. While there is a consensus that research synthesis should list inclusion/exclusion criteria, there is less discussion about how these criteria are applied in research synthesis management tools like Rayyan and EPPI Reviewer. To ensure that everyone was prepared to use these two platforms, we had a training session

> in which we practised using the platforms, compared our answers and discussed any uncertainties. However, after working independently, we found that we had used different terms to refer to the same reason (e.g., "not research synthesis" and "not synthesis"). Another issue was that we excluded articles for several reasons, and some of us labelled more reasons than others. To address these challenges, Dongxia developed a detailed workflow document for us to follow. This document specifies the terms that should be used for exclusion reasons in Rayyan, ensuring that there are no multiple labels for the same reason. The document also clarified which exclusion criterion should be evaluated first, and once this had been done, it was unnecessary to proceed to the others. Our problem-solving process highlights the benefits of more detailed planning and record-keeping in future research syntheses and other types of research involving a second reviewer or coder. For example, rather than simply having a list of inclusion/exclusion criteria, it would be more helpful to have a checklist that matches the criteria to specific actions in synthesis platforms. — Xuechun Huang

> It was my first time of using new research software to accelerate research synthesis projects, these tools are very practical and have improved my work efficiency. The first screening process was conducted on the research software Rayyan, I screened a total of 100 publications' title and abstract within two weeks. My approach was guided by the instructions provided from the meta-review project training, the predetermined inclusion criteria were the preliminary reference to my decisions on the inclusions and exclusions. I constantly checked the project guidelines to refine my selection in the first screening process. — Jing Yun

> It was easy to feel uncertain about whether to include or exclude a study. However, the team's approach to this challenge was reassuring: we had dedicated training sessions, detailed inclusion/exclusion criteria, and concrete examples to work through together. Regular meetings allowed us to go over tricky cases, and the use of Rayyan's "maybe" option gave us space to revisit ambiguous records. Over time, these structures helped me feel more confident and systematic in applying the criteria, while still recognising that judgement calls were inevitable. — Jingying Hu

To summarise, this workflow demonstrates that effective review processes were not achieved through automation alone, but through the deliberate structuring of processes that organise, calibrate, and support human judgement. Automation allowed us to work with large amounts of data and handle it in a consistent way, but the implementation and execution of the workflow ultimately depended on human decisions and actions.

2.3 Reflection on collaboration: Collective human intelligence

A third theme of our reflection is the central role of collaboration in sustaining the quality of the meta-review. As aforementioned, Applied Linguistics spans multiple subfields and methodological traditions (Pun, 2019), and no individual reviewer could fully anticipate the diversity of research encountered. The review team, composed primarily of PhD students and recently graduated PhDs with research expertise in areas such as intercultural communication, World Englishes, discourse analysis, corpus linguistics, bilingualism and multilingualism, and second language acquisition, reflected this disciplinary breadth. Collaboration allowed team members to draw on one another's disciplinary strengths, resolve uncertainties more effectively, and build a shared understanding of how the inclusion/exclusion criteria and coding scheme should be applied. In this way, individual insights accumulated into a collective human intelligence (see also Chapter 5, this volume, regarding collaborative challenges in data retrieval and keyword selection).

Collaboration also enabled iterative refinement of the workflow. Regular meetings provided opportunities to revisit ambiguous cases and adjust decision rules, while informal channels such as group chats allowed uncertainties to be addressed promptly during independent work. Timely interactions prevented small interpretive differences from becoming inconsistencies and helped reviewers develop confidence in their own decisions. Importantly, working together required negotiation and flexibility. Some decisions had no single "correct" answer, and the team often needed to balance differing perspectives before reaching a consensus. This process of co-construction ensured that the criteria were applied thoughtfully rather than mechanically, and that the project could respond to special cases without sacrificing rigour. In this sense, collaboration was not simply a logistical necessity but a methodological asset.

As Qi Liu eloquently noted, "a big team works like a factory where everyone has different roles". For our team, the process went even further: roles were dynamically adjusted according to expertise, availability, and emerging needs, so that each member could contribute meaningfully and also gain substantial learning opportunities. Below are reflections from the team members that highlight how collaboration shaped their understanding, supported their decision-making, and strengthened the project as a whole.

> Collaboration in such a large team is not easy. The commitment of team members is crucial for ensuring the quality of work. A big team works like a factory where everyone has different roles. From my perspective, overall, our team had an experienced researcher managing the overall direction, a member responsible for planning, managing, and conducting the project, and other members carrying out tasks and giving feedback. These responsibilities are not fixed; a supportive

team allowed them to shift depending on team members' availability and expertise. Communication was the key to working efficiently and sharing responsibilities. Clear action points made the research process more manageable. — Qi Liu

In terms of teamwork, our collaboration was marked by two notable strengths. First, we adhered to a protocol and screening workflow that were iteratively developed and regularly revisited. Rather than treating the protocol as static, we engaged with it as a living document — open to critical discussion and revision in light of emerging uncertainties. This flexible yet structured approach enhanced both the reliability and reflexivity of our decisions. Second, our regular team meetings served not only as checkpoints for progress but also as spaces for collaborative sense-making. A particularly valuable practice was led by Dongxia, who facilitated group analysis of articles on which screening decisions had diverged. By collectively re-examining titles, abstracts, and individual rationales, we uncovered tacit interpretive differences and refined our shared judgement criteria. These discussions often generated new insights — not just about specific studies, but about the underlying decision-making processes that shaped our synthesis. — Xin Fang

The nature of the project as a meta-review of a wide-ranging academic field required the team to screen and code research synthesis of diverse types and on numerous topics. Besides, the scale of the project demanded a large team from multidisciplinary backgrounds. The first challenge for me is to reach a shared understanding with other team members on inclusion/exclusion criteria and coding scheme. Take the screening process as an example. Although preliminary criteria were established at the beginning, they were subject to interpretations. To address this, several rounds of weekly trial-and-discussion on the same group of entries were held before we started independent screening. This stage is crucial for me to achieve a shared understanding with the team. Through trial screening, I was able to experiment with the criteria, delineating my understanding and identifying confusion. The subsequent discussion after each round of trial enabled me to clarify confusion and calibrate my understanding against others'. During the discussion, a multidisciplinary team proved to be beneficial as well, as we helped each other in understanding synthesis types and topics. Furthermore, the discussions created a dialogic space for new ideas to be shared and built on, which enhanced the rigour of the screening criteria. — Yi Zhang

Another aspect that stood out was the collaborative experience across countries and institutions. Despite differences in time zones and schedules, communication was smooth and inclusive. Online meetings were conducted regularly, and recordings were shared so members who had teaching or other commitments could catch up. Beyond meetings, our group chat created a supportive space where questions could be raised and quickly addressed. What impressed me most was the leads' patience and sense of responsibility. They frequently checked in, clarified uncertainties, and encouraged us to share difficulties. This style of coordination fostered a strong sense of belonging and accountability. Even though the

project was large, I never felt peripheral; rather, it was an immersive, participatory experience where international collaboration enriched our reflections and sharpened our decisions. For my future work, I will carry forward both sets of lessons: the methodological discipline of systematic screening and extraction, and the collaborative ethos that makes large-scale international research both rigorous and enriching. — Jingying Hu

Another good practice was maintaining a group chat to ask and answer each other's questions. We continued to use emails and online meetings for communicating key deadlines, tasks and training. However, the group chat proved to be a valuable addition. As the meta-review has a broad scope, there are many articles to screen. For a smaller team conducting synthesis, it may be feasible to discuss questions and disagreements during meetings. However, for a large team of reviewers like this one, it is more efficient to solve difficult problems in a chat so that we did not have to wait until the next meeting. —Xuechun Huang

For the data extraction stage, we shifted to EPPI Reviewer and the core team was streamlined to five members, which increased both efficiency of communication and level of responsibility. We followed the same process as in the screening stages: training, trial, discussion, another trial, and further discussion until we felt aligned enough to proceed with independent work. It is noteworthy that "independent" never meant "isolated", as our group chats remained lively: whenever someone hit a difficult case, they would bring it to the team. Instead of passively carrying out assigned tasks, each of us actively contributed to the refinement of the coding scheme by questioning ambiguous definitions and sharing our expertise. As a result, the extraction process was not mechanical but dynamic, continually updating to capture the nuances of Applied Linguistics reviews.

— Dominique Yixue Jiang

I also realised how important it is for the whole team to reach a shared understanding of the criteria. As a whole team, we had lots of discussions in either the official meetings or the informal group chats to make sure that we were all on the same page. These two channels also gave space for questions or tricky cases to be raised. For me, this is a reminder that research synthesis is not an individual task — it works best when team members collaborate and openly discuss questions.

—Rui Yan

The team was interdisciplinary, and was selected according to specific criteria and the project's needs. In spite of the time difference with other team members, I have managed to maintain work efficiency, and prioritised it over other commitments. Group communications were conducted and documented efficiently through training, WeChat, and emails. Publications were allocated at random to every team member with an agreed number of studies for the two screening stages. Each member screened studies independently and completed their own tasks following the prescheduled timeline. Any disagreements were resolved through group discussions. We constantly checked the original inclusion criteria,

> discussed case by case, and an inclusion checklist was also established to guide the screening decisions to avoid any misinterpretations. It was encouraging to see every member's aspiration toward research synthesis, and the commitment to the meta-review project. —Jing Yun

Overall, these reflections show that collaboration operated as a form of collective human intelligence, yielding conceptual clarity and methodological coherence that neither individual reviewers nor automated systems could have produced on their own.

2.4 Reflection on learning: How human reviewers grew through the process

A final theme in our reflection relates to how the reviewing process facilitated our development as researchers. In an era where AI tools can streamline many aspects of research synthesis, this project reminded us that the act of reviewing itself is a powerful learning process, one that machines cannot replicate on our behalf. Engaging with a large and diverse body of studies exposed us to unfamiliar subfields, new methodological traditions, and different ways of conceptualising Applied Linguistics. This required us to interpret varied terminology, understand distinct reporting conventions, and gradually build a more comprehensive sense of the field.

As we moved through cycles of calibration, trial screening, and protocol refinement, we began to internalise the reasoning behind research synthesis. What initially felt uncertain gradually became more intuitive. Tools such as the checklist and coding scheme helped structure our work, but it was the practice of making decisions, negotiating ambiguity, defending interpretations, and reflecting on methodological choices, that strengthened our confidence. Through this process, we learned not only how to apply a protocol, but how to understand and justify the decisions behind it. Working through complexity also proved to be instructive. Ambiguous abstracts, interdisciplinary studies, unconventional formats, and special cases required judgement that no automated tool could fully provide. Each encounter with nuance sharpened our ability to read critically and handle interpretive uncertainty. We believe these skills are essential to rigorous scholarship and transferable to future research.

Importantly, this learning unfolded collectively. Observing how others reasoned through difficult cases, exchanging questions in meetings and group chats, and contributing to the refinement of the workflow created a shared learning environment. These experiences highlight a complementary insight to our earlier reflections: the value of human reviewers lies not only in what they contribute

to the review, but also in what they gain from participating in it. AI can support efficiency, but it cannot replace the intellectual, reflective, and developmental dimensions of human engagement. Below are reflections from the team that illustrate how reviewers deepened their disciplinary knowledge, strengthened their methodological skills, and grew as researchers through the process.

> I found myself revisiting inclusion decisions, refreshing my understanding of the field's scope, and confronting the inherent diversity of applied linguistic inquiry. This initial sense of uncertainty ultimately broadened my appreciation for the field's interdisciplinarity, deepening both my disciplinary literacy and my respect for the complexity of research synthesis. — Xin Fang

> Looking back, these stages witnessed a clear progression from uncertainty to confidence, and from fragmented individual judgements to a shared interpretive framework. This project not only deepened my understanding of research synthesis and Applied Linguistics but also of how specialist tools can structure and support large-scale collaboration. It also impressed upon me the importance of methodological transparency and of documenting the rationale behind decisions, however small. Most importantly, it showed me that perseverance, empowered by collective effort, can transform an initially daunting task into a rewarding experience, an insight that will guide my future research. — Dominique Yixue Jiang

> Beyond the technical and procedural lessons, this experience also deepened my understanding of Applied Linguistics as a field. I came into the project mainly thinking about Applied Linguistics in terms of language teaching and learning as well as bilingualism, and relevant research methods like corpus linguistics or discourse analysis. But during screening, I encountered studies that linked language with other aspects such as cognition, math skills, and even medical contexts. As I screened more articles, I realised that Applied Linguistics is far broader than I had learned — it's about the role of language in many dimensions of human life. Seeing this made me appreciate more the important role of language in the world. Overall, I learned from this experience that adequate preparation, regular communication, and clear guidelines are prerequisites for successful screening. In addition, I expanded my knowledge of the field and recognised the many ways language intersects with human activity. — Rui Yan

> These activities have given me a chance to not only learn the academic vitality of the deepening understanding in Applied Linguistics, but I also gained practical experience of setting priorities and resolving contradictions through the systematic research process. The practice of using research software to conduct research synthesis studies is also part of the implications for my future research experience. The meta-review project has definitely strengthened my understanding in the field of Applied Linguistics, especially the connection with real-life situations. — Jing Yun

> Participating in the screening and data extraction of this large-scale project has provided me the opportunity to both contribute and learn. The experience has presented challenges but has also been highly rewarding. Reflecting on the challenges and insights gained from overcoming them, I noted that I have benefited from methods to ensure shared understanding, streamline workflow, and prompt timely communication when participating in the project, which I believe subsequently contributed to the robustness of the process and its results. — Yi Zhang

> This is the largest project I have participated in, not only in terms of research scope, but also in team size. The meta-review started with screening more than 10,000 articles. The initial research team had three members, which later expanded to more than ten. Although I read the protocol when I first joined, my knowledge of this review developed gradually as I became more engaged in the process. Along the way, I encountered multiple challenges and gained valuable insights.
>
> I also gained a better understanding of why a meta-review was chosen rather than other types of reviews. We intended to present research in Applied Linguistics, but the number of studies was overwhelming. Thus, we included research syntheses, each representing a group of studies in Applied Linguistics. I think this was a feasible way to capture the overall picture of research in this field. In sum, this project deepened my understanding of Applied Linguistics, taught me how to handle complex screening decisions, and highlighted the value of teamwork and communication in large-scale collaborative research. — Qi Liu

In short, these reflections show that the reviewing process was not only an analytic and synthesis task but also a meaningful site of disciplinary and methodological development. AI may streamline certain procedures, but it cannot replicate the conceptual growth, reflexive practice, and collaborative learning that human reviewers gained through the process.

3. Beyond automation: Integration of human work and AI

Across this meta-review project, we came to view the relationship between human reviewers and AI tools not as a choice between alternatives but as an evolving partnership. Automatic tools supported nearly every stage of the workflow: enhancing comprehensiveness in the search, increasing efficiency during deduplication and screening, structuring full-text review and data extraction, and improving transparency through systematic documentation. These tools strengthened the procedural foundations of the project and enabled us to manage a large and heterogeneous dataset with far greater efficiency than manual methods alone would have made possible.

At the same time, our experience consistently demonstrated the boundaries of automation (see also Chapter 12, this volume, for affordances and limitations of AI capabilities). While AI can highlight patterns, prioritise records, or generate draft extractions, it cannot interpret conceptual nuance, handle ambiguous cases, or adapt flexibly to the irregularities that inevitably arise in research reporting. The knowledge required to determine whether a study aligns with the scope of Applied Linguistics, to understand methodological claims, and to make sense of unconventional formats remained firmly human. The most intellectually demanding work (i.e. reasoning, negotiating meaning, building shared interpretations) could not be outsourced.

Our team's experience with Elicit made these boundaries even clearer. We had experimented with Elicit for screening and data extraction, and were genuinely impressed by how automatic and efficient it can be in predicting the likelihood that an article meets the inclusion criteria and in identifying key pieces of information within the text. However, these strengths quickly reveal their limits. As noted by the developer itself, while language models can support summarisation and extraction, they often miss conceptual nuance, misinterpret numerical or methodological details, and are currently unable to evaluate the epistemological quality or trustworthiness of research beyond surface-level heuristics (Elicit, 2025). Decisions about whether a study conceptually fits within the scope of Applied Linguistics, how its methodological claims should be understood, or how ambiguous reporting should be interpreted cannot be delegated to automated predictions.

In practice, we found that every AI-generated suggestion required careful verification, and in many cases, substantial correction. There is no doubt that AI tools will continue to develop and refine, and it is difficult to predict how further these tools can go. What this project has shown is that AI enhances research synthesis most effectively when paired with structured human judgement. The protocol, checklist, and coding scheme gave human reviewers a shared framework; AI tools supported efficiency and consistency within that framework. Yet it was human expertise, individually exercised and collectively developed, that guided interpretation, resolved uncertainty, and ensured conceptual coherence. Just as AI helped us manage scale, human reviewers helped the project maintain meaning.

AI use declaration statement

During the preparation of this chapter, the authors used ChatGPT (OpenAI) and Microsoft Copilot to assist with language editing in order to improve spelling and grammar accuracy and to enhance clarity and coherence. All AI-generated outputs were critically reviewed, revised,

and verified by the authors. The authors take full responsibility for the content of this manuscript.

References

Abogunrin, S., Muir, J. M., Zerbini, C., & Sarri, G. (2025). How much can we save by applying artificial intelligence in evidence synthesis? Results from a pragmatic review to quantify workload efficiencies and cost savings. *Frontiers in Pharmacology*, *16*.

Alaqlobi, O., Alduais, A., Qasem, F., & Alasmari, M. (2024). Artificial intelligence in applied (linguistics): A content analysis and future prospects. *Cogent Arts & Humanities*, *11*(1).

Chong, S. W. (2025). Synthesis Methods and Reporting Tool (SMART) for research syntheses in Applied Linguistics. *Research Synthesis in Applied Linguistics*, *1*(1), 17–38.

Chong, S. W., Nie, D., & Liu, Q. (2025). Protocol for a meta-review on Applied Linguistics researchscape. *Research Synthesis in Applied Linguistics*, *1*(2), 459–469.

Chong, S., & Plonsky, L. (2024). A typology of secondary research in Applied Linguistics. *Applied Linguistics Review*, *15*(4), 1569–1594.

Digital Solution Foundry and EPPI Centre (2023). *EPPI-Mapper* (Version 2.2.4). EPPI Centre, UCL Social Research Institute, University College London

Elicit. (2025). *Elicit: The AI research assistant* [Computer software]. https://elicit.com

Hampson, T., & McKinley, J. (2024). Systematic review, systematic bias? An example from EMI research. *Language Teaching*, 1–9.

Marshall, I. J., & Wallace, B. C. (2019). Toward systematic review automation: A practical guide to using machine learning tools in research synthesis. *Systematic Reviews*, *8*(1), 163.

Ouzzani, M., Hammady, H., Fedorowicz, Z., & Elmagarmid, A. (2016). Rayyan — A web and mobile app for systematic reviews. *Systematic Reviews*, *5*(1), 210.

Plonsky, L., Isaacs, T., & Taylor, L. (2025). Opportunities and challenges for research synthesis in Applied Linguistics: Three scholars reflect. *Research Synthesis in Applied Linguistics*, *1*(1), 5–16.

Pun, J. (2019). Interdisciplinary research. In J. McKinley & H. Rose (Eds.), *The Routledge handbook of research methods in Applied Linguistics* (pp. 108–121). Routledge.

Qureshi, R., Shaughnessy, D., Gill, K. A. R., Robinson, K. A., Li, T., & Agai, E. (2023). Are ChatGPT and large language models "the answer" to bringing us closer to systematic review automation? *Systematic Reviews*, *12*(1), 72.

Révész, A., Suzuki, S., & Jung, Y. (2025). Artificial intelligence in applied linguistics: Applications, promises, and challenges. *Annual Review of Applied Linguistics*, *45*, 1–9.

Scherbakov, D., Hubig, N., Jansari, V., Bakumenko, A., & Lenert, L. A. (2025). The emergence of large language models as tools in literature reviews: a large language model-assisted systematic review. *Journal of the American Medical Informatics Association*, *32*(6), 1071–1086.

Thomas, J., Graziosi, S., Brunton, J., Ghouze, Z., O'Driscoll, P., Bond, M., & Koryakina, A. (2023). *EPPI Reviewer: Advanced software for systematic reviews, maps and evidence synthesis* [Computer software]. EPPI Centre Software. UCL Social Research Institute, London.

CHAPTER 18

Responding to the challenges of evidence synthesis

Some lessons from healthcare

Anastasios Bastounis, Burak Kundakci & Andrew Booth
University of Sheffield

While healthcare has developed a rich ecosystem of systematic review types and standardised practice, evidence synthesis in Applied Linguistics remains comparatively nascent. Drawing on the robust methodological standards established in healthcare, the chapter considers how researchers in Applied Linguistics might benefit from adopting and adapting tools and frameworks for quality appraisal and managing heterogeneity. At the same time, it cautions against the uncritical transfer of methods, recognising that the nature of research questions, data types, and evidence use in Applied Linguistics can differ substantially. By offering practical insights and suggestions from good practice, the chapter seeks to equip Applied Linguistics researchers with foundational knowledge to enhance the rigour, transparency, and utility of their evidence syntheses, while fostering critical awareness of field-specific challenges.

Keywords: systematic reviews, evidence synthesis, Applied Linguistics, meta-analysis, quality appraisal, heterogeneity

Introduction

Systematic reviews widely regarded as the "*gold standard*" of research methodologies have significantly advanced evidence-based practice (Murad et al., 2016; Weisz et al., 2023). Renowned for their scientific transparency, methodological rigour, and reproducibility, systematic reviews constitute the cornerstone for informing medical and public health decision-making (Gurevitch et al., 2018; Munn et al., 2018; Sarri et al., 2022). The methodological rigour and the reporting of systematic reviews have been further strengthened and standardised by the introduction of the GRADE[1] approach (Guyatt et al., 2008; Oxman, 2004) and

1. https://www.gradeworkinggroup.org/

https://doi.org/10.1075/rmal.17.18bas
Available under the CC BY-NC-ND 4.0 license.© 2026 John Benjamins Publishing Company

the PRISMA Statement (Moher et al., 2009; Page et al., 2021) with its extensions.[2] Aside from conventional systematic reviews, other forms of evidence synthesis have emerged, aiming to meet the demands of a changing research landscape (Sutton et al., 2019) (See Table 1). For a complementary discussion on scoping reviews, see Chapters 12 and 13.

Table 1. Evidence synthesis alternatives (Sutton et al., 2019)

Approach	Brief description
Integrative reviews	Synthesise both experimental and non-experimental research using broad methods (Sutton et al., 2019)
Living reviews	Continually update assembled studies to incorporate new evidence in real time (Elliott et al., 2017)
Mapping and Scoping reviews	Explore research areas to guide the development of systematic reviews (Campbell et al, 2023).
Overviews of reviews (umbrella reviews)	Summarise an assembled collection of systematic reviews and meta-analyses covering a common topic (Gates et al., 2020)
Rapid reviews	Expedite evidence synthesis by streamlining or omitting certain components of conventional systematic reviews (Tricco et al., 2015)
Realist reviews	Use a theory-driven method that seeks to explore for whom, how, under what circumstances and in what context and extent complex interventions and health programmes work (Pawson et al., 2005; Wong et al., 2013)

The versatility of evidence synthesis formats has contributed to a sustained increase in production, driven in part by advancements in machine learning and natural language processing (Khalil et al., 2022). While methodological approaches to evidence syntheses remain relatively less developed within Applied Linguistics, their publication has shown a remarkable increase over the last 10 years (Amini Farsani et al., 2021; Amini Farsani & Jamali, 2023). Evidence synthesis can play a vital role in Applied Linguistics by facilitating the translation of research findings into practice (Chong et al., 2024; Plonsky, 2013). However, methodological developments specific to Applied Linguistics synthesis are a priority, given the field's idiosyncrasies (this issue is also discussed in Chapter 14) and still-nascent guidance for conduct and reporting (this issue is also discussed in Chapter 2) (Chong, 2025). Additionally, standardised reporting of evidence syntheses in Applied Linguistics is needed to strengthen reproducibility. Recent

2. https://www.equator-network.org/

efforts have examined synthesis trends and practices, focusing on topics addressed, methodologies used, and the identification of research gaps (Chalmers et al., 2024; Chong et al., 2025).

Given the relatively nascent nature of evidence synthesis in Applied Linguistics, this chapter draws upon examples from healthcare, where robust, standardised approaches to synthesising research findings are well established, to examine key methodological and ethical challenges. The chapter concludes by outlining potential courses of action for Applied Linguistics reviewers and signposting further reading, aiming to equip researchers with key lessons from evidence synthesis practice.

Methodological challenges in evidence synthesis

Key methodological challenges underpin all stages of evidence synthesis, although some more readily apply to conventional systematic reviews and meta-analyses.

Search strategy development

Search strategy development in health research has evolved significantly over recent decades. Information specialists have focussed on creating efficient search strategies that maximise relevant results while minimising irrelevant ones. To streamline the process, teams have developed reusable search blocks, as "hedges" and "filters" that are incorporated within larger search strategies. Pre-designed components include variations for specific parameters such as age groups, interventions, or contexts (hedges) or permutations of study types or publication types (filters). Search terms can be structured hierarchically, from "probable" terms for focussed searches to "possible" terms for broader coverage, and finally "safety net" terms for comprehensive retrieval. The Information Specialists' Support Group (ISSG) website maintains collections of these search examples. Peer Review of Electronic Search Strategies (PRESS) guidelines provide a comprehensive framework for evaluating search strategies, including the use of Boolean operators and database-specific subject headings (McGowan et al., 2016; Saif-Ur-Rahman, 2022). Significant improvements have occurred in study identification, particularly with MEDLINE's retrospective indexing of RCTs (Wieland et al., 2012). This has facilitated creation of specialised databases such as the Cochrane Library for RCTs and Epistemonikos for reviews (Rada et al., 2020). The field has also developed frameworks for structuring review questions (Booth et al., 2019), typologies of review questions (Munn et al., 2018) and designs (Sutton et al., 2019).

Applied Linguistics could benefit substantially from adopting rigorous, replicable approaches to study retrieval. The development of shared resources, techniques, and special interest groups[3] could foster expertise in specific Applied Linguistics areas.

Data extraction

Data extraction is a critical step in evidence synthesis, during which key information is collected from primary studies for analysis (Mathes et al., 2017). Errors during data extraction are not uncommon in healthcare research (Jones et al., 2005; Kanukula et al., 2024). Errors during this process whether caused by transcription mistakes, misinterpretation of data, or omission of critical details, can compromise the validity of findings. For instance, data extraction performed by a single reviewer is associated with a higher likelihood of errors during the extraction process (Buscemi et al., 2006). Such errors can have serious implications, including selective reporting bias, which affects meta-analysis estimates. In the absence of rigid recommendations regarding extraction (Büchter et al., 2020), a key approach to reducing errors is developing and pilot-testing of discipline-specific data extraction forms for Applied Linguistics research questions and then training data abstractors using these forms (Li et al., 2015). However, further standardisation of the data extraction process, as well as the semi-automation tools employed in the extraction, is needed to reduce the likelihood of errors (Büchter et al., 2023).

Heterogeneity and quality appraisal of evidence

Heterogeneity refers to the variability among studies included in an evidence synthesis (Glasziou & Sanders, 2002; Ioannidis et al., 2007). Heterogeneity can be broadly categorised into three types: clinical heterogeneity (Gagnier et al., 2012), methodological heterogeneity (Rhodes et al., 2018), and statistical heterogeneity (Higgins & Li, 2022). A high degree of heterogeneity, may preclude statistical pooling, making it challenging to group studies into meaningful categories and conduct direct comparisons. Where statistical pooling is possible and reviewers undertake meta-analyses, the presence of heterogeneity should be quantified (e.g., I^2 statistic) and potential sources of heterogeneity should be investigated using either subgroups analyses or meta-regressions (Higgins & Li, 2022).

3. https://researchsynthesis.weebly.com/

Quality appraisal involves evaluating the risk of bias in included studies; however, this process is not always straightforward. The terms "*quality assessment*" and "*risk of bias assessment*" have been used interchangeably in literature, even though they have distinct meanings. Quality assessment evaluates whether a study includes methodological safeguards to minimise bias, whereas risk of bias assessment considers how the presence or absence of such safeguards influences study results (Furuya-Kanamori et al., 2021). Quality assessment is thus defined as the extent to which a study is designed, conducted, analysed, interpreted, and reported to avoid systematic errors, while risk of bias assessment specifically examines how methodological flaws impact study outcomes (Banzi et al., 2018).

It is not uncommon for evidence syntheses and systematic reviews to include both RCTs and observational or non-randomised studies. Reviewers then use multiple quality appraisal tools to accommodate different study designs. Notwithstanding apparent similarities, each tool evaluates different categories, making direct comparisons of methodological quality across studies with different designs challenging. Additionally, different tools are available for assessing the same study designs. While these tools often overlap, they are not identical, adding further complexity.

The GRADE framework extends the concept of quality appraisal beyond individual study assessments (Schünemann et al., 2013). In GRADE, the term "*quality of evidence*" reflects confidence in the adequacy of an effect estimate to support a particular recommendation. This evaluation goes beyond risk of bias in individual studies to consider four additional factors: inconsistency, indirectness, imprecision, and publication bias (See Balshem et al., 2011; Guyatt et al., 2011). Another major challenge in quality appraisal is its reliance not only on *methodological* quality but also on the *reporting* quality of the studies. Studies with suboptimal reporting may receive lower ratings, even if their methods are robust.

Non-reporting bias and statistical challenges

Evidence synthesis faces further challenges from the threat of bias in the primary studies included in evidence syntheses. While non-reporting bias is most pertinent to systematic reviews and meta-analyses of RCTs, it can affect all types of evidence synthesis. Non-reporting bias can lead to inflated effect sizes, threatening the validity of systematic review findings (Kirkham et al., 2010). In this subsection, we focus on key aspects of selective (non-)reporting bias within systematic reviews and meta-analyses (See Boutron et al., 2024; Chan & Altman, 2005; Song et al., 2010). Although examples are predominantly drawn from healthcare, it is highly likely that certain types of bias extend to Applied Linguistics (Table 2).

Reviewers should thus consider the possibility that such types of bias are similarly present in evidence syntheses and systematic reviews within Applied Linguistics.

Table 2. Type of bias frequently encountered in systematic reviews

Types of bias	Description
Selective (non-) reporting bias	Typically occurs when studies report only a subset of measured outcomes — particularly those that are statistically significant or favourable — while omitting or underreporting non-significant or negative results
Publication bias	Typically occurs when studies with statistically significant results are more likely to be published than those with non-statistically significant results.
Grey literature bias	Typically occurs when results from unpublished studies (e.g., governmental/agency reports, dissertations or conference abstracts) systematically differ compared to results reported in peer-reviewed journal articles
Time-lag bias	Typically occurs when studies with statistically significant results are published more quickly than those with non-significant results.
Language and database bias	Typically occurs when studies published in English and indexed in major databases are prioritised, potentially excluding relevant studies published in other languages or formats.
Outcome reporting bias	Typically occurs when there is the selective reporting of some of the outcomes (plus incomplete reporting of outcomes) and/or when not all analysed outcomes are reported.

Reviewers often encounter methodological challenges related to statistical procedures required for conducting a meta-analysis. We therefore highlight key aspects for reviewers conducting meta-analyses to consider carefully (See Deeks et al., 2024): (a) selecting the appropriate meta-analytic model (e.g., fixed- or random-effects model) according to the contextual and methodological characteristics of the included studies, (b) addressing and interpreting the observed heterogeneity, (c) handling missing data in the included primary studies, (d) defining the rationale for subgroup and sensitivity analyses, explicitly stating the proposed moderators of action, (e) aggregating different effect sizes for the same outcome (e.g., OR, RR, SMD) and assessing whether necessary transformations are feasible based on the data provided in the included studies, (f) combining multiple intervention arms in order to render them comparable in conventional meta-analysis, and (g) considering, in the context of primary studies with mul-

tiple arms and where sufficient data exists, the multiple comparison of interventions simultaneously by combining direct and indirect evidence across studies (i.e., network meta-analysis) (Chaimani et al., 2024). Network meta-analysis combines direct and indirect estimates across a network of interventions in a single analysis and can inform comparative effectiveness of multiple interventions even when direct head-to-head comparisons don't exist (See Ades et al., 2024; Chaimani et al., 2024; Dias et al., 2018; White, 2015). For a complementary discussion on meta-analyses, see Chapters 7 and 8.

Mixed-methods reviews, reviews of complex interventions and overviews of reviews

Mixed-methods synthesis presents additional challenges for reviewers due to methodological variations across included primary studies, stemming from differences in data types, analytical approaches, and interpretive frameworks (Hong et al., 2017). One major challenge involves how to integrate incompatible data types, as quantitative data provide numerical estimates, while qualitative data yield themes and patterns. Reviewers must decide, from the outset, how they will synthesise the data. This includes considering the chosen synthesis framework (e.g., explanatory/exploratory sequential), accounting for heterogeneity among included studies, and determining how to prioritise or weigh evidence from different data types.

Mixed-methods reviews are particularly appropriate when evaluating complex interventions (Moore et al., 2015). Complex interventions often involve multiple components delivered in varying contexts (Craig et al., 2008; Skivington et al., 2021), making it challenging to disentangle causal mechanisms and assess their effectiveness through conventional systematic reviews. Key methodological challenges arise when assessing complex interventions within evidence syntheses, such as: (a) variability in intervention components of complex interventions, (b) contextual dependence as the effectiveness of complex interventions frequently depends on contextual factors, (c) outcome heterogeneity, (d) limited process evaluations, and (e) theoretical frameworks, as complex interventions are often based on integrated theoretical frameworks, requiring substantial expertise within the evidence synthesis team. A theory-based review approach, the realist synthesis, is designed specifically to understand complex interventions. Such a review seeks to explore what works, for whom, in what circumstances, and why, by exploring the underpinning mechanisms through which interventions produce outcomes, and the contexts within which these mechanisms are activated (Wong et al., 2017).

Mixed-methods synthesis (Tricco et al., 2016) may be well-suited to the Applied Linguistics context, where researchers might combine experimental studies with qualitative research. Careful selection of appropriate synthesis designs can help highlight the underpinning mechanisms of complex and culturally sensitive interventions. Methodological frameworks like GRADE-CERQual (Lewin et al., 2018) can further support transparent assessment of confidence in qualitative findings, making them particularly valuable for informing policy.

Overviews of reviews (i.e., "umbrella" reviews) summarise findings from systematic reviews, where the unit of searching, inclusion, and data analysis is the ***systematic review itself*** rather than the individual primary study (Aromataris et al., 2015; Pollock et al., 2024). These overviews use systematic and explicit methods, similar to conventional systematic reviews, to synthesise data from multiple systematic reviews on specific research topics (Pollock et al., 2024). Increased numbers of systematic reviews in healthcare and high levels of duplication across published reviews present methodological challenges for reviewers conducting overviews of reviews (Gates et a., 2020). Methodological challenges can be summarised as follows (Gates et al., 2020): (a) high level of redundancy and duplication amongst systematic reviews, (b) the existence of multiple systematic reviews increases the likelihood of conflicting conclusions, (c) overlapping primary studies included in the systematic reviews may lead to double-counting of data, (d) overviews often include systematic reviews with varying levels of methodological quality, (e) when multiple systematic reviews exist within the same field, it is difficult for reviewers to decide which reviews to include in their overview, (f) when the scope of the overview of reviews is not identical to the scope of included systematic reviews, reviewers might need to search for individual studies reported within the included systematic reviews or even supplementary primary studies. Overviews of reviews are well-suited for synthesising evidence from a large and fragmented body of evidence syntheses. Recent efforts have begun to explore synthesis methodologies and identify research gaps in Applied Linguistics (Chong et al., 2025). Using validated appraisal tools (e.g., AMSTAR-2, ROBIS) and overlap metrics to quantify duplication across included primary studies can help mitigate bias and enhance transparency. Overviews of reviews could expedite the methodological development of synthesis within Applied Linguistics, helping to prioritise future research directions while reducing redundancy and research waste.

Ethical challenges in evidence synthesis

Key ethical challenges in conducting and reporting evidence syntheses include ensuring inclusivity and representation, transparently reporting methodological decisions, and making defensible trade-offs between comprehensiveness and feasibility. Reviewers must also use unpublished data ethically and remain reflexive to minimise confirmation bias in interpretation.

Transparency, power imbalances and ethical use of data

Reviewers should clearly document the methods, data sources, eligibility criteria, and analytical or interpretative approaches to maintain methodological rigour and ensure reproducibility. Closely tied to transparency is the presence of conflicts of interest (CoI) in evidence syntheses. Failure to appropriately address CoI can introduce bias into the analysis and interpretation of results. This, in turn, may distort research priorities, favour third-party interests, and undermine the ethical objectives of the research (Bekelman et al., 2003; Bero, 2017). It is interesting to note within healthcare how emphasis has switched from ***avoiding*** conflicts of interest, increasingly challenging within an interconnected research community, to ***identifying and acknowledging*** potential conflicts and handling them proportionately. A review team also needs to ensure transparency through accurate attribution of authorship and appropriate acknowledgment of contributions (Allen et al., 2014). Unethical practices, such as ghost authorship, defined as the non-inclusion as an author of individuals who have contributed substantially to the research of an article (Dotson & Slaughter, 2011), and honorary authorship, the inclusion of an individual who does not meet basic authorship criteria (ICMJE, 2025; Rennie & Flanagin, 1994), undermine research integrity, compromising the credibility of the research process (Bates et al., 2004). Further ethical challenges related to power imbalances include plagiarism and redundancy. While not exclusive to evidence synthesis, plagiarism (including self-plagiarism) and redundancy (e.g., salami slicing), defined as the inappropriate and deliberate splitting of research data in order to achieve greater numbers of minor publications (COPE, 2023), represent serious ethical violations that undermine research integrity and credibility. Similarly, selective citation practices exert detrimental effects on the validity of synthesis findings.

Implications for practice

This section identifies practical solutions to address key methodological and ethical challenges related to the conduct of syntheses.

Search strategy development and data extraction

Optimal searches should ensure sufficient database coverage by including several major subject databases and at least one multi-disciplinary database, as well as Google Scholar (Bramer et al., 2017). Typically, searches should include sources of unpublished data, while language or date range restrictions in searches should be avoided. Reviewers are advised to conform to the PRESS guideline for developing electronic search strategies (McGowan et al., 2016), while also adhering to the PRISMA-S extension statement when reporting their literature searches (Rethlefsen et al., 2021). Independent extraction by two reviewers is considered the most rigorous option, and when relevant data is not reported in the text, primary study authors should be contacted for additional information. In cases where data are presented graphically but not in the text, appropriate software[4] should be used to extract the data directly from the graphs.

Heterogeneity and quality appraisal of evidence

When conducting a meta-analysis, reviewers should assess the degree of inconsistency (i.e., I^2 statistic). Subgroup analyses or meta-regressions should be performed to investigate likely sources of heterogeneity. In meta-analyses of RCTs, reviewers should avoid incorporating non-randomised studies into the analysis. Independent quality appraisal and risk of bias assessment by two reviewers is considered optimal. Reviewers should use validated tools that match the study designs of included studies. Where applicable, the GRADE approach should be applied to assess the certainty of the evidence.

Non-reporting bias and statistical challenges

Reviewers should seek to register review protocols prospectively in platforms like PROSPERO[5] or the Open Science Framework (OSF).[6] Additionally, reviewers

4. https://plotdigitizer.com/

5. https://www.crd.york.ac.uk/prospero/

6. https://osf.io/

should use available reporting checklists in their reviews and assess whether the included studies conform to any appropriate reporting checklist. For systematic reviews with meta-analysis, reviewers should formally investigate publication bias using appropriate statistical tests. Reviewers should clearly specify their chosen model of the meta-analysis in the protocol and provide a rationale for their decision. They should also investigate potential sources of heterogeneity, quantifying such heterogeneity using appropriate statistical measures. In addition, reviewers should indicate in their protocol how they will conduct sensitivity analyses and explicitly state the moderators they will test in subgroup analyses.

Mixed-methods reviews and overviews of reviews

Reviewers should describe in detail in the protocol how qualitative and quantitative data will be integrated, utilising explicit frameworks. They should consider using the Mixed Methods Appraisal Tool (MMAT) (Hong et al., 2018) to assess the methodological quality of mixed-methods studies. Additionally, GRADE or GRADE-CERQual tools should be considered to assess the confidence of evidence in quantitative and qualitative studies, respectively (Noyes et al., 2019). Reviewers should also consider applying framework or thematic synthesis and assess the implementation fidelity of included interventions (See Wiltsey Stirman et al., 2019).

Reviewers should seek to register or publish the protocol for their overview in advance and adhere to reporting checklists (i.e., PRIOR statement) (Gates et al., 2022). The use of validated tools such as AMSTAR-2 (Shea et al., 2017) or ROBIS (Whiting et al., 2016) for assessing methodological quality and risk of bias is strongly encouraged. From the outset the review team should decide whether and how to include overlapping systematic reviews (see Pollock et al., 2024). They should seek to assess and quantify the overlap among primary studies included in the systematic reviews using appropriate statistical software (Bougioukas et al., 2023).

Transparency and reproducibility

Reviewers should declare any CoI, ensure that their evidence synthesis undergoes independent peer review, and, where possible, involve diverse stakeholders. Where feasible, they should apply frameworks such as PROGRESS-Plus (O'Neill et al., 2014) to evaluate equity dimensions in the conduct and reporting of evidence synthesis. Additionally, reviewers should adhere to the ICMJE criteria for

authorship (ICMJE, 2025) and implement tools such as CRediT[7] to specify individual contributions. Reviewers should adhere to COPE (Committee on Publication Ethics) guidelines for ethical publishing practices. These include disclosing any prior publications related to the review, justifying overlaps and explicitly stating what their evidence synthesis or systematic review adds to the prevalent literature.

Conclusion

Within healthcare, evidence synthesis has benefited from positioning the evidence base as central to decision-making, the sheer volume of activity and endeavour, coordination at national and international levels and early involvement in seeking to resolve methodological issues. Rather than simply repeating such an evolution, Applied Linguistics should seek to absorb and assimilate earlier lessons from the healthcare field.

Further reading

Booth, A., Sutton, A., Clowes, M., & Martyn-St James, M. (2026). *Systematic approaches to a successful literature review* (4th ed.). Sage.

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2021). *Introduction to meta-analysis*. John Wiley & Sons.

A foundational textbook offering a clear and accessible introduction to the principles and methods of meta-analysis.

Higgins, J. P. T., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M. J., & Welch, V. A. (Eds.). (2024). *Cochrane handbook for systematic reviews of interventions* (Version 6.5, updated August 2024). Cochrane. https://training.cochrane.org/handbook

Egger, M., Higgins, J. P. T., & Smith, G. D. (2022). *Systematic reviews in health research: Meta-analysis in context* (3rd ed.). John Wiley & Sons.

AI use declaration statement

The authors confirm that no AI tools were used in the preparation, writing, formatting, and editing of this work.

7. https://credit.niso.org/

References

Ades, A. E., Welton, N. J., Dias, S., Phillippo, D. M., & Caldwell, D. M. (2024). Twenty years of network meta-analysis: Continuing controversies and recent developments. *Research Synthesis Methods*, *15*(5), 702–727.

Allen, L., Brand, A., Scott, J., Altman, M., & Hlava, M. (2014). Credit where credit is due. *Nature*, *508*, 312–313.

Amini Farsani, M., & Jamali, H. R. (2023). Collaboration network of Applied Linguistics research articles with different methodological orientations. *Studies in Second Language Learning and Teaching*, *13*(4), 727–754.

Amini Farsani, M., Jamali, H. R., Beikmohammadi, M., Ghorbani, B. D., & Soleimani, L. (2021). Methodological orientations, academic citations, and scientific collaboration in Applied Linguistics: What do research synthesis and bibliometrics indicate? *System*, *100*.

Aromataris, E., Fernandez, R., Godfrey, C. M., Holly, C., Khalil, H., & Tungpunkom, P. (2015). Summarizing systematic reviews: Methodological development, conduct and reporting of an umbrella review approach. *International Journal of Evidence Based Healthcare*, *13*(3), 132–140.

Balshem, H., Helfand, M., Schunemann, H. J., Oxman, A. D., Kunz, R., Brozek, J., Vist, G. E., Falck-Ytter, Y., Meerpohl, J., Norris, S., & Guyatt, G. H. (2011). GRADE guidelines: 3. Rating the quality of evidence. *Journal of Clinical Epidemiology*, *64*(4), 401–406.

Banzi, R., Cinquini, M., Gonzalez-Lorenzo, M., Pecoraro, V., Capobussi, M., & Minozzi, S. (2018). Quality assessment versus risk of bias in systematic reviews: AMSTAR and ROBIS had similar reliability but differed in their construct and applicability. *Journal of Clinical Epidemiology*, *99*, 24–32.

Bates, T., Anic, A., Marusic, M., & Marusic, A. (2004). Authorship criteria and disclosure of contributions comparison of 3 general medical journals with different author contribution forms. *Jama*, *292*, 86–88.

Bekelman, J. E., Li, Y., & Gross, C. P. (2003). Scope and impact of financial conflicts of interest in biomedical research: A systematic review. *Jama*, *289*(4), 454–465.

Bero, L. (2017). Addressing bias and conflict of interest among biomedical researchers. *Jama*, *317*, 1723–1724.

Booth, A., Noyes, J., Flemming, K., Moore, G., Tuncalp, O., & Shakibazadeh, E. (2019). Formulating questions to explore complex interventions within qualitative evidence synthesis. *BMJ Glob Health*, *4*(Suppl. 1), e001107.

Booth, A., James, M. S., Clowes, M., & Sutton, A. (2025). *Systematic approaches to a successful literature review* (4th ed.). Sage.

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2021). *Introduction to meta-analysis*. John Wiley & sons.

Bougioukas, K. I., Diakonidis, T., Mavromanoli, A. C., & Haidich, A. B. (2023). ccaR: A package for assessing primary study overlap across systematic reviews in overviews. *Research Synthesis Methods*, *14*(3), 443–454.

Boutron, I., Page, M. J., Higgins, J. P. T., Altman, D. G., Lundh, A., Hróbjartsson, A., & Group, C. B. M. (2024). Chapter 7: Considering bias and conflicts of interest among the included studies. In J. P. T. Higgins, J. Thomas, J. Chandler, M. Cumpston, T. Li, M. J. Page, & W. VA (Eds.), *Cochrane handbook for systematic reviews of interventions* (Version 6.5.). Cochrane.

doi Bramer, W. M., Rethlefsen, M. L., Kleijnen, J., & Franco, O. H. (2017). Optimal database combinations for literature searches in systematic reviews: A prospective exploratory study. *Systematic Reviews*, *6*(1), 245.

doi Büchter, R. B., Rombey, T., Mathes, T., Khalil, H., Lunny, C., Pollock, D., Puljak, L., Tricco, A. C., & Pieper, D. (2023). Systematic reviewers used various approaches to data extraction and expressed several research needs: A survey. *Journal of Clinical Epidemiology*, *159*, 214–224.

doi Büchter, R. B., Weise, A., & Pieper, D. (2020). Development, testing and use of data extraction forms in systematic reviews: A review of methodological guidance. *BMC Medical Research Methodology*, *20*.

doi Buscemi, N., Hartling, L., Vandermeer, B., Tjosvold, L., & Klassen, T. P. (2006). Single data extraction generated more errors than double data extraction in systematic reviews. *Journal of Clinical Epidemiology*, *59*(7), 697–703.

doi Campbell, F., Tricco, A. C., Munn, Z., Pollock, D., Saran, A., Sutton, A., White, H., & Khalil, H. (2023). Mapping reviews, scoping reviews, and evidence and gap maps (EGMs): The same but different- the "Big Picture" review family. *Systematic Reviews*, *12*(1):45.

Chaimani, A., Caldwell, D. M., Li, T., Higgins, J. P. T., & Salanti, G. (2024). Chapter 11: Undertaking network meta-analyses. In J. P. T. Higgins, J. Thomas, J. Chandler, M. Cumpston, T. Li, M. J. Page, & W. VA (Eds.), *Cochrane handbook for systematic reviews of interventions* (Version 6.5.). Cochrane.

doi Chalmers, H., Brown, J., & Koryakina, A. (2024). Topics, publication patterns, and reporting quality in systematic reviews in language education. Lessons from the international database of education systematic reviews (IDESR). *Applied Linguistics Review*, *15*(4), 1645–1669.

doi Chan, A. W., & Altman, D. G. (2005). Identifying outcome reporting bias in randomised trials on PubMed: Review of publications and survey of authors. *BMJ*, *330*(7494), 753.

doi Chong, S. W. (2025). Synthesis Methods and Reporting Tool (SMART) for research syntheses in Applied Linguistics. *Research Synthesis in Applied Linguistics*, *1*(1), 17–38.

doi Chong, S. W., Bond, M., & Chalmers, H. (2024). Opening the methodological black box of research synthesis in language education: Where are we now and where are we heading? *Applied Linguistics Review*, *15*(4), 1557–1568.

doi Chong, S. W., Nie, D., & Liu, Q. (2025). Protocol for a meta-review on Applied Linguistics researchscape. *Research Synthesis in Applied Linguistics*, *1*(2), 459–469.

doi Craig, P., Dieppe, P., Macintyre, S., Michie, S., Nazareth, I., Petticrew, M., & Medical Research Council, G. (2008). Developing and evaluating complex interventions: The new Medical Research Council guidance. *BMJ*, *337*, a1655.

Deeks, J. J., Higgins, J. P. T., Altman, D. G., McKenzie, J. E., & Veroniki, A. A. (2024). Chapter 10: Analysing data and undertaking meta-analyses. In J. P. T. Higgins, J. Thomas, J. Chandler, M. Cumpston, T. Li, M. J. Page, & W. VA (Eds.), *Cochrane handbook for systematic reviews of interventions* (Version 6.5.). Cochrane.

Dias, S., Ades, A. E., Welton, N. J., Jansen, J. P., & Sutton, A. J. (2018). *Network meta-analysis for decision-making*. John Wiley & Sons.

Dotson, B., & Slaughter, R. L. (2011). Prevalence of articles with honorary and ghost authors in three pharmacy journals. *American Journal of Health System Pharmacy*, *68*(18), 1730–1734.

Egger, M., Higgins, J. P. T., & Smith, G. D. (2022). *Systematic reviews in health research: Meta-analysis in context* (3rd ed., pp. 185–203). John Wiley & Sons.

Elliott, J. H., Synnot, A., Turner, T., Simmonds, M., Akl, E. A., McDonald, S., Salanti, G., Meerpohl, J., MacLehose, H., Hilton, J., Tovey, D., Shemilt, I., Thomas, J., on behalf of the Living Systematic Review Network. (2017). Living systematic review: 1. Introduction-the why, what, when, and how. *Journal of Clinical Epidemiology*, *91*, 23–30.

Furuya-Kanamori, L., Xu, C., Hasan, S. S., & Doi, S. A. (2021). Quality versus risk-of-bias assessment in clinical research. *Journal of Clinical Epidemiology*, *129*, 172–175.

Gagnier, J. J., Moher, D., Boon, H., Beyene, J., & Bombardier, C. (2012). Investigating clinical heterogeneity in systematic reviews: A methodologic review of guidance in the literature. *BMC Medical Research Methodology*, *12*.

Gates, M., Gates, A., Guitard, S., Pollock, M., & Hartling, L. (2020). Guidance for overviews of reviews continues to accumulate, but important challenges remain: a scoping review. *Systematic Reviews*, *9*(1), 254.

Gates, M., Gates, A., Pieper, D., Fernandes, R. M., Tricco, A. C., Moher, D., Brennan, S. E., Li, T., Pollock, M., Lunny, C., Sepulveda, D., McKenzie, J. E., Scott, S. D., Robinson, K. A., Matthias, K., Bougioukas, K. I., Fusar-Poli, P., Whiting, P., Moss, S. J., & Hartling, L. (2022). Reporting guideline for overviews of reviews of healthcare interventions: Development of the PRIOR statement. *BMJ*, *378*, e070849.

Glasziou, P. P., & Sanders, S. L. (2002). Investigating causes of heterogeneity in systematic reviews. *Statistics in Medicine*, *21*(11), 1503–1511.

Gurevitch, J., Koricheva, J., Nakagawa, S., & Stewart, G. (2018). Meta-analysis and the science of research synthesis. *Nature* (7695), 175–182.

Guyatt, G. H., Oxman, A. D., Kunz, R., Brozek, J., Alonso-Coello, P., Rind, D., Devereaux, P. J., Montori, V. M., Freyschuss, B., Vist, G., Jaeschke, R., Williams, J. W., Jr., Murad, M. H., Sinclair, D., Falck-Ytter, Y., Meerpohl, J., Whittington, C., Thorlund, K., Andrews, J., & Schunemann, H. J. (2011). GRADE guidelines 6. Rating the quality of evidence — Imprecision. *Journal of Clinical Epidemiology*, *64*(12), 1283–1293.

Guyatt, G. H., Oxman, A. D., Kunz, R., Woodcock, J., Brozek, J., Helfand, M., Alonso-Coello, P., Falck-Ytter, Y., Jaeschke, R., Vist, G., Akl, E. A., Post, P. N., Norris, S., Meerpohl, J., Shukla, V. K., Nasser, M., Schunemann, H. J., & Group, G. W. (2011). GRADE guidelines: 8. Rating the quality of evidence — Indirectness. *Journal of Clinical Epidemiology*, *64*(12), 1303–1310.

Guyatt, G. H., Oxman, A. D., Kunz, R., Woodcock, J., Brozek, J., Helfand, M., Alonso-Coello, P., Glasziou, P., Jaeschke, R., Akl, E. A., Norris, S., Vist, G., Dahm, P., Shukla, V. K., Higgins, J., Falck-Ytter, Y., Schunemann, H. J., & Group, G. W. (2011). GRADE guidelines: 7. Rating the quality of evidence — Inconsistency. *Journal of Clinical Epidemiology*, *64*(12), 1294–1302.

Guyatt, G. H., Oxman, A. D., Sultan, S., Glasziou, P., Akl, E. A., Alonso-Coello, P., Atkins, D., Kunz, R., Brozek, J., Montori, V., Jaeschke, R., Rind, D., Dahm, P., Meerpohl, J., Vist, G., Berliner, E., Norris, S., Falck-Ytter, Y., Murad, M. H., Schunemann, H. J., & Group, G. W. (2011). GRADE guidelines: 9. Rating up the quality of evidence. *Journal of Clinical Epidemiology*, *64*(12), 1311–1316.

Guyatt, G. H., Oxman, A. D., Vist, G., Kunz, R., Brozek, J., Alonso-Coello, P., Montori, V., Akl, E. A., Djulbegovic, B., Falck-Ytter, Y., Norris, S. L., Williams, J. W., Jr., Atkins, D., Meerpohl, J., & Schunemann, H. J. (2011). GRADE guidelines: 4. Rating the quality of evidence — Study limitations (risk of bias). *Journal of Clinical Epidemiology*, *64*(4), 407–415.

Guyatt, G. H., Oxman, A. D., Vist, G. E., Kunz, R., Falck-Ytter, Y., Alonso-Coello, P., & Schünemann, H. J. (2008). GRADE: An emerging consensus on rating quality of evidence and strength of recommendations. *BMJ*, *336*.

Higgins, J. P. T., & Li, T. (2022). Exploring heterogeneity. In M. Egger, J. P. T. Higgins, & G. D. Smith (Eds.), *Systematic reviews in health research: Meta-analysis in context* (3rd ed., pp. 185–203). John Wiley & Sons.

Hong, Q. N., Pluye, P., Fàbregues, S., Bartlett, G., Boardman, F., Cargo, M., Dagenais, P., Gagnon, M.-P., Griffiths, F., Nicolau, B., O'Cathain, A., Rousseau, M.-C., Vedel, I., & Pluye, P. (2018). *Mixed methods appraisal tool (MMAT), version 2018: User guide [PDF]*. McGill University. https://mixedmethodsappraisaltoolpublic.pbworks.com/

Hong, Q. N., Pluye, P., Bujold, M., & Wassef, M. (2017). Convergent and sequential synthesis designs: Implications for conducting and reporting systematic reviews of qualitative and quantitative evidence. *Systematic Reviews*, *6*(1), 61.

ICMJE. (2025). Recommendations for the conduct, reporting, editing, and publication of scholarly work in medical journals. http://www.icmje.org/recommendations/

Ioannidis, J., Patsopoulos, N., & Evangelou, E. (2007). Uncertainty in heterogeneity estimates in meta-analyses. *BMJ*, *335*, 914–918.

Jones, A. P., Remmington, T., Williamson, P. R., Ashby, D., & Smyth, R. L. (2005). High prevalence but low impact of data extraction and reporting errors were found in Cochrane systematic reviews. *Journal of Clinical Epidemiology*, *58*(7).

Kanukula, R., Page, M. J., Turner, S. L., & McKenzie, J. E. (2024). Identification of application and interpretation errors that can occur in pairwise meta-analyses in systematic reviews of interventions: a systematic review. *Journal of Clinical Epidemiology*, *170*, 111331.

Khalil, H., Ameen, D., & Zarnegar, A. (2022). Tools to support the automation of systematic reviews: a scoping review. *Journal of Clinical Epidemiology*, *144*, 22–42.

Kirkham, J. J., Dwan, K. M., Altman, D. G., Gamble, C., Dodd, S., Smyth, R., & Williamson, P. R. (2010). The impact of outcome reporting bias in randomised controlled trials on a cohort of systematic reviews. *BMJ*, *340*, c365.

Lewin, S., Booth, A., Glenton, C., Munthe-Kaas, H., Rashidian, A., Wainwright, M., Bohren, M.-A., Tunçalp, Ö., Colvin, C.-J., Garside, R., Carlsen, B., Langlois, E.-V., & Noyes, J. (2018). Applying GRADE-CERQual to qualitative evidence synthesis findings: Introduction to the series. *Implementation Science*, *13*(Suppl-1), Article-2.

Li, T., Vedula, S. S., Hadar, N., Parkin, C., Lau, J., & Dickersin, K. (2015). Innovations in data collection, management, and archiving for systematic reviews. *Annals of Internal Medicine*, *162*(4), 287–294.

Mathes, T., Klaßen, P., & Pieper, D. (2017). Frequency of data extraction errors and methods to increase data extraction quality: a methodological review. *BMC medical research methodology*, *17*(1), 152.

McGowan, J., Sampson, M., Salzwedel, D. M., Cogo, E., Foerster, V., & Lefebvre, C. (2016). PRESS peer review of electronic search strategies: 2015 guideline statement. *Journal of Clinical Epidemiology*, *75*, 40–46.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & Group, P. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Medicine*, *6*(7), e1000097.

Moore, G. F., Audrey, S., Barker, M., Bond, L., Bonell, C., Hardeman, W., Moore, L., O'Cathain, A., Tinati, T., Wight, D., & Baird, J. (2015). Process evaluation of complex interventions: Medical Research Council guidance. *BMJ*, *350*, 1258.

Munn, Z., Stern, C., Aromataris, E., Lockwood, C., & Jordan, Z. (2018). What kind of systematic review should I conduct? A proposed typology and guidance for systematic reviewers in the medical and health sciences. *BMC Medical Research Methodology*, *18*(1), 5.

Murad, M. H., Asi, N., Alsawas, M., & Alahdab, F. (2016). New evidence pyramid. *BMJ Evidence-Based Medicine*, *21*(4), 125–127.

Murad, M. H., & Wang, Z. (2017). Guidelines for reporting meta-epidemiological methodology research. *Evidence Based Medicine*, *22*(4), 139–142.

Noyes, J., Booth, A., Moore, G., Flemming, K., Tuncalp, O., & Shakibazadeh, E. (2019). Synthesising quantitative and qualitative evidence to inform guidelines on complex interventions: Clarifying the purposes, designs and outlining some methods. *BMJ Global Health*, *4*(Suppl 1), e000893.

O'Neill, J., Tabish, H., Welch, V., Petticrew, M., Pottie, K., Clarke, M., Evans, T., Pardo Pardo, J., Waters, E., White, H., & Tugwell, P. (2014). Applying an equity lens to interventions: Using PROGRESS ensures consideration of socially stratifying factors to illuminate inequities in health. *Journal of Clinical Epidemiology*, *67*(1), 56–64.

Oxman, A. D. (2004). Grading quality of evidence and strength of recommendations. *BMJ*, *328*, 1490–1494.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hrobjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., McGuinness, L. A., Stewart, L. A., Thomas, J., Tricco, A. C., Welch, V. A., Whiting, P., & Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, *372*, n71.

Pawson, R., Greenhalgh, T., Harvey, G., & Walshe, K. (2005). Realist review — A new method of systematic review designed for complex policy interventions. *Journal of Health Services Research & Policy*, *10*, 21–34.

Plonsky, L. (2013). Study quality in SLA. *Studies in Second Language Acquisition*, *35*(4), 655–687.

Pollock, M., Fernandes, R. M., Becker, L. A., Pieper, D., & Hartling, L. (2024). Chapter V: Overviews of reviews. In J. P. T. Higgins, J. Thomas, J. Chandler, M. Cumpston, T. Li, M. J. Page, & W. VA (Eds.), *Cochrane handbook for systematic reviews of interventions* (Version 6.5.). Cochrane.

Rada, G., Perez, D., Araya-Quintanilla, F., Avila, C., Bravo-Soto, G., Bravo-Jeria, R., Canepa, A., Capurro, D., Castro-Gutierrez, V., Contreras, V., Edwards, J., Faundez, J., Garrido, D., Jimenez, M., Llovet, V., Lobos, D., Madrid, F., Morel-Marambio, M., Mendoza, A., Neumann, I., Ortiz-Munoz, L., Pena, J., Perez, M., Pesce, F., Rain, C., Rivera, S., Sepulveda, J., Soto, M., Valverde, F., Vasquez, J., Verdugo-Paiva, F., Vergara, C., Zavala, C., Zilleruelo-Ramos, R., & Epistemonikos, P. (2020). Epistemonikos: A comprehensive database of systematic reviews for health decision-making. *BMC Medical Research Methodology*, *20*(1), 286.

Rennie, D., & Flanagin, A. (1994). Authorship! Authorship! Guests, ghosts, grafters, and the two-sided coin. *Jama*, *271*(6).

Rethlefsen, M. L., Kirtley, S., Waffenschmidt, S., Ayala, A. P., Moher, D., Page, M. J., Koffel, J. B., & Group, P.-S. (2021). PRISMA-S: An extension to the PRISMA statement for reporting literature searches in systematic reviews. *Systematic Reviews*, *10*(1), 39.

Saif-Ur-Rahman, K. M. (2022). Transparency of reporting search strategies in systematic reviews. *Hypertension Research*, *45*(11), 1838–1839.

Sarri, G., Patorno, E., Yuan, H., Guo, J., Bennett, D., Wen, X., Zullo, A. R., Largent, J., Panaccio, M., Gokhale, M., Moga, D. C., Ali, S. M., & Debray, T. P. A. (2022). Framework for the synthesis of non- randomised studies and randomised controlled trials: A guidance on conducting a systematic review and meta- analysis for healthcare decision making. *BMJ Evidence-Based Medicine 27*, 109–119.

Schünemann, H., Brożek, J., Guyatt, G., & Oxman, A. (2013). *GRADE handbook for grading quality of evidence and strength of recommendations* (H. Schünemann, J. Brożek, G. Guyatt, & A. Oxman, Eds.). GRADE Working Group.

Shea, B. J., Reeves, B. C., Wells, G., Thuku, M., Hamel, C., Moran, J., Moher, D., Tugwell, P., Welch, V., Kristjansson, E., & Henry, D. A. (2017). AMSTAR 2: A critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. *BMJ*, *358*, j4008.

Skivington, K., Matthews, L., Simpson, S. A., Craig, P., Baird, J., Blazeby, J. M., Boyd, K. A., Craig, N., French, D. P., McIntosh, E., Petticrew, M., Rycroft-Malone, J., White, M., & Moore, L. (2021). A new framework for developing and evaluating complex interventions: Update of Medical Research Council guidance. *BMJ*, *374*, n2061.

Song, F., Parekh, S., Hooper, L., Loke, Y. K., Ryder, J., Sutton, A. J., Hing, C., Kwok, C. S., Pang, C., & Harvey, I. (2010). Dissemination and publication of research findings: An updated review of related biases. *Health Technology Assessment*, *14*(8), iii, ix–xi, 1–193.

Sutton, A., Clowes, M., Preston, L., & Booth, A. (2019). Meeting the review family: Exploring review types and associated information retrieval requirements. *Health Information and Libraries Journal*, *36*(3), 202–222.

Tricco, A. C., Antony, J., Soobiah, C., Kastner, M., MacDonald, H., Cogo, E., Lillie, E., Tran, J., & Straus, S. E. (2016). Knowledge synthesis methods for integrating qualitative and quantitative data: A scoping review reveals poor operationalization of the methodological steps. *Journal of Clinical Epidemiology*, *73*, 29–35.

Tricco, A. C., Antony, J., Zarin, W., Strifler, L., Ghassemi, M., Ivory, J., Perrier, L., Hutton, B., Moher, D., & Straus, S. E. (2015). A scoping review of rapid review methods. *BMC Medicine*, *13*, 224.

Weisz, J. R., Venturo-Conerly, K. E., Fitzpatrick, O. M., Frederick, J. A., & Ng, M. Y. (2023). What four decades of meta-analysis have taught us about youth psychotherapy and the science of research synthesis. *Annual Review of Clinical Psychology*, *19*, 79–105.

White, I. R. (2015). Network meta-analysis. *Stata Journal*, *15*(4), 951–985.

Whiting, P., Savovic, J., Higgins, J. P., Caldwell, D. M., Reeves, B. C., Shea, B., Davies, P., Kleijnen, J., Churchill, R., & group, R. (2016). ROBIS: A new tool to assess risk of bias in systematic reviews was developed. *Journal of Clinical Epidemiology*, *69*, 225–234.

Wieland, L. S., Robinson, K. A., & Dickersin, K. (2012). Understanding why evidence from randomised clinical trials may not be retrieved from Medline: Comparison of indexed and non-indexed records. *BMJ*, *344*, d7501.

Wiltsey Stirman, S., Baumann, A. A., & Miller, C. J. (2019). The FRAME: An expanded framework for reporting adaptations and modifications to evidence-based interventions. *Implementation Science*, *14*(1), 58.

Wong, G., Greenhalgh, T., Westhorp, G., Buckingham, J., & Pawson, R. (2013). RAMESES publication standards: Realist syntheses. *BMC Medicine*, *11*, 21.

Wong, G., Westhorp, G., Greenhalgh, J., Manzano, A., Jagosh, J., & Greenhalgh, T. (2017). Quality and reporting standards, resources, training materials and information for realist evaluation: The RAMESES II project. *Health Services and Delivery Research*, *5*(28).

SECTION 9

Research synthesis in Applied Linguistics

Looking ahead

CHAPTER 19

A principled approach to methodological agility in research synthesis

Putting ethics at the centre

Sin Wang Chong
University of St Andrews

Editing this book has been a learning journey for me. I feel privileged to learn from the vast and diverse experiences of the contributors. At the same time, I am impressed by how much we have evolved as a field in terms of methodological diversity and innovation for research syntheses, compared to 20 years ago. As mentioned in the introductory chapter (Chapter 1), this book is not a "how-to" guide for conducting different forms of research synthesis (although there are some mini-guides throughout the book); instead, it invites contributors and readers to consider ethical and methodological challenges and dilemmas, and how they are navigated. Reading the reflective narratives in this book does not provide *the* answer to methodological issues in research syntheses, but rather offers insight into how authors responded to these challenges, having considered their areas of research, contexts, and intentions for conducting a research synthesis. I begin this chapter by revisiting the major methodological and ethical challenges shared by authors and their responses. Then, I introduce a framework of "ethical methodological agility" and outline an ethics-first approach that can guide our response to the plethora of challenges associated with research synthesis methodologies.

Keywords: research synthesis, research ethics, methodological agility, ethical challenges

Navigating challenges

Throughout the book, contributors describe a shared set of challenges that arise when conducting and publishing research syntheses in Applied Linguistics. These challenges differ in form, but they often stem from the same sources. The chapters also document how researchers responded to these challenges and dilemmas in practice.

https://doi.org/10.1075/rmal.17.19cho
Available under the CC BY-NC-ND 4.0 license.© 2026 John Benjamins Publishing Company

Challenge 1: Some concepts are hard to define

A recurring challenge is about definition. Contributors of this book struggled with determining what constitutes a research synthesis and distinguishing between different review types. Ambiguities noted in the chapters pertain to concepts and constructs, moderator labels, inclusion criteria, and the scope of Applied Linguistics.

In response, strategies adopted by authors in this book are pragmatic. Journal editors set clear expectations through guidelines and reporting requirements. Researchers treated protocols as a "live" documents, refining definitions as the review projects progress. When areas such as language, context, or geographic scope remain ambiguous, authors were transparent about their own interpretations and trade-offs, and provide justifications about feasibility, ethics, and purpose.

Challenge 2: Methodological guidance does not always fit

Chapters in this book noted that existing research synthesis guidelines and tools were either underdeveloped for Applied Linguistics or imported from other fields. This created tensions around search strategies, appraisal frameworks, analysis.

Contributors responded by adapting methods. Applied linguists developed field-specific frameworks and coding manuals, conducted calibration exercises and inter-coder reliability checks. For instance, when conducting meta-analyses, coding was refined iteratively, referring to theories when selecting moderators, and conducted sensitivity and bias analyses. When tools from other disciplines were used, they were revised to align with research norms in Applied Linguistics, with limitations clearly acknowledged.

Challenge 3: Volume of research makes the scope of a research synthesis unmanageable

Several chapters highlight that it is not always possible to conduct research syntheses comprehensively. Conducting broad searches requires time, labour, and expertise that are not affordable by all. This is particularly true for student-led projects and diverse teams.

Teams in this book responded through careful and strategic calibration and workflow design, including conducting multiple rounds of pilot screening, partial dual screening, and clearly assigning team members to specific roles. These helped manage workload while preserving methodological rigour. For some

teams, it may mean that scopes need to be narrowed, databases limited, or review types adjusted to ensure that the review remained meaningful and manageable.

Challenge 4: Data quality is mixed and information is missing

Contributors reflected on issues related to missing metadata, inconsistent abstracts, unreliable outcome reporting, and poorly described primary studies. Technical challenges documented in the chapters relate to exporting from databases, deduplicating records, and identifying compatible software. These are addressed through manual data cleaning and checking and keeping a thorough audit trail. When key information is missing, researchers explore alternative analytical techniques and look for supplementary materials; some teams also needed to switch tools in the middle of a project or used a combination of software.

Challenge 5: It is difficult to strike a balance between quality and inclusivity

The chapters in this book discuss various forms of bias, exclusion of minority contexts, and the balance between quality and being comprehensive. Contributors responded by being reflexive in reporting and explicit with critiquing their own decisions. Some broadened their searches to include regional or non-English sources while others opted for documenting exclusions carefully and reporting on constraints such as limited database access.

Challenge 6: Collaborations are rarely straightforward

Several chapters emphasise the human side of synthesis work, including uncertainty during decision-making, stress on student researchers, and tensions between supervisors and supervisees. Such tensions can only be resolved through.collaborative supervision models, clear recordkeeping, and open dialogue, which helped distribute responsibility and reduce individual strain.

Ethical methodological agility

These six challenges show that there is no single rule of thumb that will provide a quick fix to these methodological issues. Some would argue that following the recommended steps in methodological guidelines such as PRISMA and SMART provides an authoritative response. However, with a closer examination of the operational and pragmatic side of conducting a research synthesis, there are issues that these methodological guides cannot resolve. For example, well-

established guidelines require multiple reviewers to be involved in the screening, data extraction, study appraisal, and data synthesis stages, and researchers are expected to report inter-coder reliability. As shown in some chapters, this is not possible because the synthesis is a supervised piece of work. Here is another example: it is preferred for research syntheses to be as comprehensive as possible, covering all relevant literature. This, nevertheless, may not be a realistic goal for some researchers due to time and resource constraints. Similarly, searching on multiple databases is recommended to ensure thorough coverage of a research synthesis. This poses a challenge to researchers who work in low-resource settings with limited access to subscription-based databases.

The examples outlined are some of the blockers that make researchers hesitant to conduct a research synthesis. In my experience editing and reviewing research syntheses for academic journals, researchers, when prompted, are usually transparent about the reasons behind the limitations of their synthesis. I always encourage researchers to be forthcoming about the deviations from methodological recommendations and provide a justification. This is an opportunity for researchers to be reflexive about their methodological choices, which is a crucial characteristic in rigorous research syntheses, as I argued in Chong (2025). In the same paper, I introduced four quality criteria of good research syntheses: systematicity, transparency, accessibility, and reflexivity, or STAR in short. "Systematicity" and "reflexivity" should not be viewed as contradictory. In fact, I would argue that researcher reflexivity makes research syntheses more systematic by constantly examining one's own methodological choices in light of recommended practices and identifying areas where well-justified deviations are needed. While the STAR framework provides a good starting point to think about the need to balance methodological standardisation and innovation in research syntheses, the justifications provided by authors are not always convincing to editors and reviewers.

Methodological agility

Here, I propose a framework of ethical methodological agility (Figure 1) that could inform researchers' responses to methodological challenges in research syntheses. I will first unpack what "methodological agility" is and how to approach it in a principled way by putting research ethics at the centre of methodological considerations.

Methodological agility is a research approach that intentionally integrates multiple research methods, methodologies, epistemological frameworks, and disciplinary perspectives to address complex research questions. It recognises that the problems we study often "demand synergy in methodology, holism in ontol-

ogy, pluralism/open-mindedness in epistemology, and embracing diversity" (Sui & DeLyser, 2012, p. 111). Methodological agility advocates that research designs need to be fit-for-purpose and contextualised. Focusing on the importance of agility (or hybridity) in fieldwork, Watkins (2020) contends that an agile research approach "embraces and incorporates uncertainties, challenges, and opportunities to expedite self-reflexive, situated, and flexible research designs" (p. 11). Methodological diversity is not a weakness to be managed; it is a resource for understanding phenomena from different angles.

Methodological agility is not the same as a mixed methods research approach because it extends beyond combining research techniques. It encompasses multiple ways of conceptualising research problems and knowing, especially when responding to complex interdisciplinary issues like those we tackle in Applied Linguistics. Nightingale (2016) describes such inquiry as "epistemological pluralism" (p. 41). This means approaching research problems "through more than one way of conceptualising" them (Nightingale, 2016, p. 41). Importantly, epistemological pluralism accepts that different methods may yield various, and sometimes conflicting, insights into the same phenomena. Divergence between datasets and ways these datasets are analysed can reveal "silences" (i.e., knowledge gaps) that themselves become objects of analysis (Nightingale, 2016, p. 42). As Sui and DeLyser (2012) note, methodological agility intends to "challenge existing boundaries and forge creative connections" across analytical divides (p. 113). In research synthesis, epistemological pluralism recognises that while syntheses aim to be comprehensive by presenting the state-of-play of a research topic or field, the synthesis itself only presents a partial and situated view of knowledge. This view is guided, and at times, limited, by the synthesis's research questions and the ways publications are selected and analysed.

In short, methodological agility comprises three characteristics: epistemological pluralism, interdisciplinary integration, and methodological flexibility (Figure 1). These characteristics allow researchers to remain open to "different and divergent methods" that can "tackle issues from different angles" (Sui & DeLyser, 2014, p. 303). Epistemological pluralism can be seen when a research synthesis includes both statistical evidence and qualitative accounts. As an example, a synthesis on second language motivation may report test scores of students while also including interview data that show how students talk about their feelings toward learning a new language. While the foci of these two forms of evidence differ, they can complement each other by showing gaps, tensions, or questions.

Interdisciplinary integration becomes visible when a research synthesis draws on ideas or evidence from more than one field. For example, a synthetic study on digital language learning can include studies from Applied Linguistics, Computer Science, and Psychology. Each field adds a different understanding. Linguis-

tics can explain language development, technology studies can explain how tools shape learning, and psychology can show how attention and memory change in online spaces. Looking across these fields allows the synthesis to address the topic more fully.

Finally, methodological flexibility is reflected in how the synthesis is carried out. Instead of following one fixed guideline, researchers choose what fits the question and the evidence. For example, if some studies test classroom tools while others explore student experiences, the researcher may use different analysis approaches for each. Flexibility here is not a lack of structure; it is a response to the nature of the evidence and the complexity of the questions.

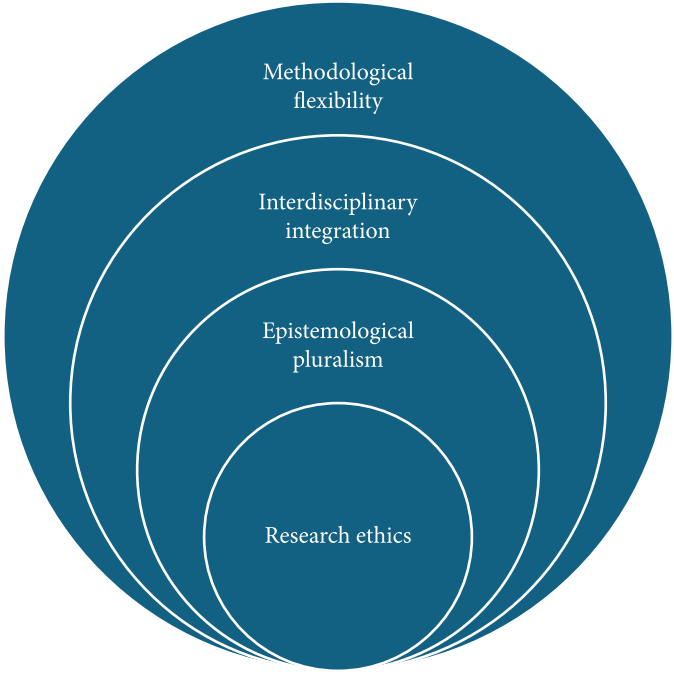


Figure 1. A framework of ethical methodological agility

Ethics before methodology

When reflecting on the literature on agility in research methodologies, one issue remains unaddressed — if we were encouraged to draw on a plethora of traditions and resources when designing a piece of research, how do researchers make decisions about their epistemological stance, methodological frameworks, and disciplinary practices? Is there a universal guiding principle that enables researchers to make informed and defensible methodological choices? To me, that anchor is

research ethics (see also Chapter 1 of this book). A lot has been written about research ethics in various disciplines such as in Education (e.g., BERA, 2024); within Applied Linguistics, contemporary perspectives on research ethics centre around three considerations: respect for people (including researchers and participants), maximising benefits and reducing harm, and ensuring justice (De Costa et al., 2021; De Costa et al., 2024). Ethics operates at two levels: macroethics addresses procedural requirements from research ethics committees, while microethics concerns the everyday decisions researchers make throughout their projects (Yaw et al., 2023). These day-to-day choices carry ethical weight because they shape how research unfolds. Scholars argue that research ethics must account for contextual complexities. From an emic perspective, ethical decision-making becomes an ongoing process of assessment and reassessment as issues arise (BERA, 2024). Doing so requires reflexivity; for example, researchers are encouraged to examine their own biases and power relations with participants (De Costa et al., 2021).

Methodological rigour is itself an ethical obligation. Poor methods not only limit the benefits of a synthesis but can mislead future research (Isbell & De Costa, 2024; Plonsky, 2024). Equally, transparency, the honest and transparent reporting of methodological procedures, is a precondition for others to assess and build on the work. Ethics also encompasses questions of societal value; for instance, research should contribute meaningfully beyond academia (De Costa et al., 2021). This broader view positions research ethics as a foundation for quality scholarship, including for researchers to approach methodological agility when conducting research syntheses.

When applying methodological agility in research synthesis, ethics should shape how decisions are made. In other words, ethical methodological agility answers the question of how others can benefit from this review. In research synthesis, "doing good" has several layers. It means doing good to the researchers whose studies are included; to researchers who will use the synthesis in their own work; to researchers who conduct the synthesis, especially when they face limits in time, access, or training. It also extends to practitioners and policymakers who rely on the synthesis to make decisions. Finally, it means doing good to the field by producing knowledge that is trusted, honest, and useful.

Doing good to researchers whose studies are included

Doing good to primary researchers means treating their work with respect. Ethical questions include: *Have I represented their findings accurately? Have I avoided oversimplifying their results?* For example, if a study reports mixed outcomes, the synthesis should not only highlight the positive trend because it reads better. It

should present the complexity that the primary researcher documented. It also means citing studies fairly, not using selective quotes, and not dismissing work from less well-known contexts.

Doing good to researchers who use the synthesis

Researchers may rely on research syntheses when developing new research projects and locating gaps in the evidence. Important ethical questions such as Am I explicit about what this synthesis cannot address? Am I honest about missing information? need to be considered. If a synthesis only considers English-language publications, it should note that work in other languages may show different patterns. This helps future researchers avoid repeating the same blind spots. Ethical decisions are made to avoid misdirection, making sure the synthesis supports, rather than misguides, future work.

Doing good to researchers who conduct the synthesis

Synthesists themselves must be protected from unrealistic methodological demands. Ethical questions such as *Am I being honest about my unique challenges? Am I naming the barriers instead of hiding them?* need to be asked. In situations where a synthesist cannot access paid databases, an ethical response is to state this openly, reflect on its impact, report alternative solutions, and describe what might be missing from the analysis. It is important to be transparent about contextual constraints.

Doing good to users of the synthesis

Teachers, professionals, and policymakers make decisions based on research syntheses. Synthesists must consider the following questions: *Will my claims guide practice based on available evidence? Am I making exaggerated claims?* When intervention studies show improvement in one context but not another, the synthesis should not claim universal effectiveness. Conclusions in research syntheses must be supported by robust evidence. Findings must be written in language that practitioners and policymakers can understand.

Doing good to the field of research

Doing good to the field means contributing knowledge that is trusted, transparent, and meaningful. Ethical questions to ask are: *Does this synthesis leave the field clearer, not more confused? Am I adding insight rather than noise?* Instead of

repeating common claims, a research synthesis might highlight under-researched groups or question long-held assumptions. Research synthesis should strengthen how the field understands itself and support collective growth as a research community.

The ethical dimensions above, while not exhaustive, provide some points of reference when we strike a balance between standardisation in methodological practices and their innovation. When ethics is given priority, an agile methodological design in research synthesis becomes an ethical practice of care: for researchers and the knowledge we build together.

This book has shown that there is no single way to conduct a research synthesis in Applied Linguistics. There is also no perfect synthesis. When ethics guides our methodological choices, our work becomes more careful, more transparent, and more trustworthy. As I noted in the opening chapter, 2026 is a special year for Applied Linguistics in relation to research synthesis. Over the past twenty years, we have seen rapid growth in this area, along with creative uses of methods drawn from other disciplines. Looking to the next twenty years, I hope our focus turns to doing research synthesis well, not just doing more of it. An ethics-first approach can guide us toward work that is rigorous and meaningful for everyone who relies on it, including researchers, practitioners, and the field as a whole.

AI use declaration statement

This chapter was written entirely by the author. After the draft was completed, Claude Sonnet 4.6 (Anthropic) was used for copyediting. Specifically, it was used to check for internal inconsistencies and contradictions. All observations generated by the AI tool were reviewed and verified by the author before any changes were made. The author takes full responsibility for the content of this chapter.

References

British Educational Research Association. (2024). *Ethical guidelines for educational research* (5th ed.). British Educational Research Association.

Chong, S. W. (2025). Synthesis Methods and Reporting Tool (SMART) for research syntheses in Applied Linguistics. *Research Synthesis in Applied Linguistics*, *1*(1), 17–38.

DeLyser, D., & Sui, D. (2014). Crossing the qualitative-quantitative chasm III: Enduring methods, open geography, participatory research, and the fourth paradigm: Enduring methods, open geography, participatory research, and the fourth paradigm. *Progress in Human Geography*, 38(2), 294–307.

De Costa, P. I., Rabie-Ahmed, A., & Cinaglia, C. (2024). *Ethical issues in applied linguistics scholarship*. John Benjamins. https://benjamins.com/catalog/rmal.7?srsltid=AfmBOopn6l1gGFGlgrHnO9gzYtt4HgdcxQ2GhIu4DpurJNADVQOCtCYK

De Costa, P. I., Sterling, S., Lee, J., Li, W., & Rawal, H. (2021). Research tasks on ethics in applied linguistics. *Language Teaching*, *54*(1), 58–70.

Isbell, D. R., & De Costa, P. I. (2024). Ethics in applied linguistics research: Evidence and action. *Research Methods in Applied Linguistics*, *3*(1), 100104.

Nightingale, A. J. (2016). Adaptive scholarship and situated knowledges? Hybrid methodologies and plural epistemologies in climate change adaptation research. *Area*, *48*, 41–47.

Plonsky, L. (2024). Study quality as an intellectual and ethical imperative: A proposed framework. *Annual Review of Applied Linguistics*, *44*, 4–18.

Sui, D., & DeLyser, D. (2012). Crossing the qualitative-quantitative chasm I: Hybrid geographies, the spatial turn, and volunteered geographic information (VGI). *Progress in Human Geography*, *36*(1), 111–124.

Watkins, C. (2020). The field and the work: Hybridity as mantra and method*. *Geographical Review*, *110*(1–2), 8–22.

Yaw, K., Plonsky, L., Larsson, T., Sterling, S., & Kytö, M. (2023). Research ethics in Applied Linguistics. *Language Teaching*, *56*(4), 478–494.

Index

A
abstracts 204, 232, 238–239, 243
accuracy 232, 333–334
adjusted Hedges' g 123
AI tools 329, 332–333, 335, 343, 345–346
AI-assisted tools 329, 332
AI-driven tools 331
AI-supported tools 329, 333
AI 6, 10, 18–19
Aims 203
AMSTAR-2 355, 358
Analysis 204, 234, 237–238, 244, 246
Anglophone 179, 181–185, 187–189, 191–194, 196
Anglosphere *See* Anglophone 183
anticipated 75–77
Applied Linguistics 4, 17–18, 71, 73–74, 78, 86–89, 136, 149, 151, 201, 203–206, 228, 230–231, 233–234, 237, 240–243, 245–246, 347, 349–351, 353, 355, 359–361, 367, 369, 372–374, 376–377
appraisal 293, 298–299, 301, 304–306, 308–309
Article 204
artificial intelligence 240, 244, 246
AXIS 250

B
bias 8, 10, 49–51, 247, 251, 352–353, 355–356, 361
beliefs 293, 297, 306
bibliographic data 67, 70, 91, 94–95, 99, 105, 109
bibliographic software 239
bibliometric 26, 40, 44, 51, 53, 55–56, 71, 73–81, 84–89
Bibliometric 26, 71, 89, 201
book chapters 74–76
Book Reviews 60
books 74–76, 78
Boolean 350

C
concept model 117, 120, 124, 128
challenges 231, 235, 237, 239–240, 243–244, 246
checklist 335–339, 343, 346
citation chaining 69, 71, 91, 249
citation patterns 71, 73, 88
clarity 333, 338, 343
classroom intervention 184–185, 190
classroom 230, 235, 238–239, 243, 245
Cochrane 350, 359, 361–363, 365
coding procedures 151
coding scheme 336–337, 340–343, 346
Coding 43, 46, 48, 50–51, 135–144, 146–147, 150–151, 201, 205, 232, 238–240, 243
collaboration 231, 245, 329, 331, 340–344
collaborative 179, 195
collective reflections 331
complex interventions 354, 360–361, 364–365
comprehensive review 31, 35
comprehensiveness 332, 345
Concept 203, 205, 234, 243
Conclusion 204
conflicts of interest 356, 360–361
COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) Risk of Bias Checklist 28
consistency 333, 335, 337, 346
Content and Language Integrated Learning 235
Context 203, 234, 243
corpus linguistics 71, 73, 76–77, 81–82, 84–85, 88
Covidence 250
criteria 203–206
Critical appraisal 260, 262–263
Curriculum developers 203

D
Data extraction 8–10, 204, 214, 219, 222, 224, 239, 244, 331, 336–337, 342, 345–346, 351
data mining 236
data synthesis 8–9
database search 273–274, 283
Database 203
Databases 203–204, 228, 232–234, 236–240, 243–245, 257, 259
delayed posttest 117, 119, 124, 127
dependencies 114, 116, 131
design 291, 293, 296, 298–300, 304, 307
differential item functioning (DIF) 28, 35
dilemmas 232, 234, 236, 241–242, 245
doctoral 291, 294, 297–298, 306

E
EAL 252, 254–259, 261, 263, 265–270
Editorials 203
Educational level 203
Educational Resources Information Center (ERIC) 29
effect sizes 46, 49, 136, 139, 147–148, 151–153
effectiveness 293, 297
effects 297, 299, 301
efficiency 329, 334–335, 337–339, 342, 344–346
eligibility 203, 239, 249
EMEMUS 183, 193
EMI 183–184, 193
empirical studies 230, 240
Empirical 203

English Medium Education in Multilingual University Settings *See* EMEMUS 183
English Medium of Instruction *See* EMI 183
English 230, 232–236, 240, 243–247
epistemological 243
EPPI Reviewer 250, 332, 334, 336, 338–339, 342
equity, diversity and inclusion 240
ERIC 203
Ethical consideration 75, 78–79, 83, 87, 136–138, 140, 142, 234, 238, 240–241, 245, 294–295, 297–298, 302, 304, 307, 370, 374–375, 377
ethical methodological agility 367, 371, 373
Ethical 75, 78–79, 83, 87, 136–138, 140, 142, 234, 238, 240–241, 245, 294–295, 297–298, 302, 304, 307, 370, 374–375, 377
ethics 1, 9–10, 15–16, 367, 369, 374, 376–377
evidence synthesis 4, 16, 18–19, 25, 347, 349–352, 354, 356, 358–360, 363
Evidence 201, 203–204, 206
exclusion criteria 203, 228, 256, 259
See inclusion and exclusion criteria
Exclusion criteria 203, 256, 259, 266
exclusion 71, 73, 75, 77, 79–81, 83
Exploratory 201, 203
Extraction of data 203

F
Findings 201, 204, 206
flow diagram Framework 201, 203, 205–206
Full text 204

G
generalizability 44, 50, 53
Geographical settings 203
Gorard's Sieve 28, 250
GRADE-CERQual 355, 358
GRADE 349, 352, 355, 357–358, 360, 362–363, 365
Grey literature 119, 203, 353
guidelines 32

H
HE 182–183, 185, 189, 191, 193–194
heterogeneity 347, 351, 353–354, 357–358, 362–363
Higher Education *See* HE 182, 228, 231, 235–238, 241, 245
human judgement 329, 332, 335, 339, 346
human reviewers 329, 332–333, 335, 344–346
human-led, AI-facilitated 215, 222–223

I
I^2 statistic 351
ICC 122–123, 127
Identification 203
Inclusion 71, 73–78, 80–81, 83–84, 87
Inclusion criteria 203, 256, 258, 261, 264–265, 331, 339, 343, 346
inclusion/exclusion criteria 203, 334, 336, 338–341
indexing database 75
influential 71, 73–76, 85
Infographics 63
innovation 235, 237
institutional support 241
intact-class adjustments 123
inter-coder reliability 8, 241
Intercoder reliability 201
International Database of Education Systematic Reviews 28, 230, 245
(IDESR) 29
interpretation of the data 242–243
Intervention 252, 254, 256–257, 259–261, 264, 269, 291, 293, 303, 306, 309
issues 228, 231–235, 237, 240, 242

J
Joanna Briggs Institute 250
journal articles 74–76, 78, 231, 235, 237, 244
journal Keywords 73, 75, 77, 79, 81–82, 85, 203

K
knowledge user engagement 238, 241

L
L2 assessment 35
L2 proficiency 34
language assessment 1, 3, 7–10, 12, 14–15, 17–19, 25, 31
Language Behavior Abstracts 203
language bias 270, 277–278, 282, 290
language interpreting 35
language proficiency test 30, 35
language teachers 234
language teaching 231, 234, 246–247
language testing 23, 25, 27–28, 35
languages 203, 228, 230–231, 233–237, 240, 243, 245
Learners 203, 205
learning 232, 245–246, 338, 340, 343, 345
Literature review 203
LLBA *See* Language Behavior Abstracts 203

M
(M)UTOS framework 116, 118, 120, 124, 127, 130
Mapping reviews 201
meaningful 295–296, 298, 301, 306
Meta-analysis 26, 30–31, 34–35, 40, 42, 44, 46–47, 51–56, 111, 133, 135, 137, 139–145, 147–154, 228, 230, 246, 351, 353, 357–360, 362
meta-regressions 352, 357
meta-review 329, 331–333, 335–336, 338–345
metadata 75–76, 78–79, 82–85, 238
Methodological agility 71, 73, 75, 77–80, 83, 87–88, 201, 204–206, 228, 232, 240–241, 243–245, 291, 293–295, 299, 302, 304–305, 307, 367, 370–372, 374

Methodological Articles 59
methodological decision-making 151
methodological framework 89, 92, 97–98, 105–106
methodological quality 27–28
Methodological review 26, 310
methodological robustness 329, 333–334
Methodological Stages 111
methodological synthesis 26, 45, 54, 56
methodological 71, 73, 75, 77–80, 83, 87–88, 201, 204–206, 228, 232, 240–241, 243–245, 291, 293–295, 299, 302, 304–305, 307
Methodology 228, 237, 244
missing information 239
Mixed Methods Appraisal Tool *See* MMAT 250, 294
Mixed-methods 203, 354–355, 358
MMAT 294, 304–306, 309, 358, 363
moderator inclusion 125
moderator 136–139, 142–144, 147–148, 150–151
Multilingual 64, 181, 183–184, 198
multivariate multilevel meta-analysis 114, 116, 129, 131–132

N
narrative analysis 228, 231
narrative review 31
narrative synthesis 250–252
note-taking 114, 116–122, 124–126, 128–129, 131–133

O
OASIS 237
Objective. *See* Aim objectives. *See* aims 203–204
open access 25, 35
Open Science Framework (OSF) 28, 242
open science 27, 114, 130
open-access 56, 64–65
Opinion pieces 203
Overviews of reviews 355

P
Parents 203
participants 234
pedagogical practices 241
percentages 116–117
performance analysis 69, 93
planning 233, 238
Policymakers 203
Population Concept, Context (PCC) framework 203, 234, 243
Population positionality 242
practice review 155, 157–160, 179, 184–192, 194–196, 198
Practices 201, 205
practitioners 179, 182, 184, 186, 190, 192, 194–195, 197
pre-registration 28–29, 61
Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) 24, 26–27, 32
PRESS 350, 357, 364
primary research 23, 25, 30
PRISMA 32, 69, 71, 100, 203, 207, 298, 300–301, 303, 305, 307, 310, 349, 357, 364–365
professional knowledge 239
PROSPERO 28–29, 242
protocol registration 61
Protocols 60
proxy 122, 125, 127
publication bias 28, 136, 139, 141–142, 144, 146, 151, 352, 358
Publication languages 203
publication trend 24, 29
publications 114

Q
Q methodology 310
Q-sort 312
qualitative research synthesis (QRS) 155
qualitative synthesis of research (QSR) 155
Qualitative 201, 203, 206
quality 293–295, 298–300, 302–305, 307–308
quality appraisal 250, 347, 351–352, 357
Quantitative 201, 203
question 247, 249
questions 203–205, 249

R
Rapid reviews 349
Rasch measurement 35
rater effect 35
rater-mediated rubric 28
rating quality 35
rating scale development 35
rationale 203, 206
Rayyan 250, 331–332, 334–336, 338–339
Realist reviews 349
reflection 179, 185, 191–195
reflections *See* reflection
reflexivity 9, 241, 278, 288, 371, 374
registered report 24, 26, 28–29
relevance 331–333, 338
Reliability 40, 48, 51, 53–56, 73, 75, 78, 81, 83, 201, 206
replication study 23
replication 40, 42, 51, 53, 55
reporting 204, 296, 299–301, 303–304, 307–310
reporting bias 351–353, 357, 361, 363
reporting quality 24, 26–27
reporting standards 26, 31
reproducibility 124, 232, 347, 350, 356, 358
research 201, 203–207, 228, 230–243, 245–246
Research ethics 367, 371, 374
research protocol 28
research syntheses 367, 369, 371, 375
Research Synthesis 1, 4–10, 12–17, 59, 136, 150–151, 201, 228
research waste 295, 298–300, 307, 309
Researchers 201, 204, 228, 230–232, 235–236, 238, 240–244
Results 204
reviewers 331, 350–351, 353–355, 357–359, 361, 364
rigor 43, 51
rigour 236, 240–241, 244, 294, 300
risk of bias 352, 357–358, 360, 363, 366
ROBIS 355, 358, 360, 366
RS. *See* Research synthesis 204

S
science mapping 69, 92, 107
scientometric 71, 87–89
Scoping review *See* ScR 26–28, 31, 35, 201, 203, 205–207, 211, 213, 228
scoping studies 201
Scopus 71, 73–76, 78–82, 84, 86–89, 203
scoring rubric 35
ScR, *See* scoping review 201, 203–204, 228, 230–234, 237–238, 240–244
screening 6, 8, 10, 14, 19, 203–206, 213, 218, 252, 257–261, 263–265, 267, 270, 331–339, 341–346
Screening criteria 332, 335, 341
Search 203
approach. *See* search strategy 203
Search strategy 203–204, 206, 234–235, 238, 249, 251
Search terms 255, 263–264, 266
Search secondary research 24–27, 29, 35, 232, 237
Selection 203, 205
self-assessment 30, 34
sensitivity analyses 353, 358
shared understanding 335, 337, 340–342, 345
SLA 43–45, 52–56
SMART 34, 62, 184, 196–197, 204, 288, 290
See Synthesis Methods and Reporting Tool (SMART) 34
SMART software tools 113
Spanish 230, 233, 244, 246
Speakers 203
speech evaluation 35
SSLA 40, 42–51, 54
stakeholder consultation 215, 225
Stakeholders 201, 238, 242–243
state-of-the-art review 26, 31
statistic synthesis 247, 250–251
statistical 71, 73–74, 82, 85–86
strategies 230–231, 234, 244
study appraisal 8, 10
Subgroup analyses 357
Synthesis Methods and Reporting Tool (SMART) 206, 228, 231, 236, 238, 293–294, 297–298, 301, 303–304, 308
Synthesis Methods and Reporting Tool *See* SMART 204
synthesis 206, 228, 231, 236, 238, 293–294, 297–298, 301, 303–304, 308
systematic examination 31, 73
Systematic review 24, 26–28, 30–32, 35, 136, 139, 145, 150, 152–153, 201, 247, 250–252, 270, 272–279, 281–282, 284–291, 347, 352, 355, 359–360, 362–365
systematic 73
systematicity 332

T
task effect 228
Teacher trainers 203
Teachers 203
teachers' professional development 232
Team 201, 204
Teaching, *See* TT
team teaching 228, 230–231, 235–237, 239
TESOLgraphics 64
test component ordering 28
test performance 34
test preparation 35
testing phase 235
Theoretical 203, 205
time period 203
timeframe 235
Title registration 61
Title Topics 201, 203, 205
translanguaging 179, 181–187, 189–196, 198–199
transparency 42–43, 50, 135–136, 138–139, 142, 147–148, 332, 334, 336, 344–345, 347, 355–356
transparent 293–295, 299, 301, 304–305
TT, *See* team teaching 231–233, 235–239, 242

U
umbrella reviews 27, 349
unanticipated 75, 79
unethical 296, 299–300

V
validity 40, 45, 53, 55, 75, 81–82
verification 238
visualization 70, 96, 102
vocabulary 133, 136–137, 141–142, 153–154

W
wasteful 296–297, 299–300
WCF 73–77, 79, 88
Web of Science 203
well-being 136, 140, 144–145, 149–150
workflow 329, 331–332, 335–341, 343, 345

This edited volume offers a critical examination of research synthesis in Applied Linguistics at a time when such work is increasingly influential and increasingly misunderstood. The volume brings together reflective accounts of conducting and editing different types of research syntheses. Instead of a procedural manual, the book foregrounds the ethical and methodological decisions that shape research syntheses in practice. Across the book, contributors discuss opportunities, constraints, and tensions between standardisation and innovation. The book is written for researchers, postgraduate students, journal editors, and methodologists who are interested in research synthesis as a reflexive, situated, and collaborative form of scholarship.

"Characterized by methodological rigor, transparent reporting, comprehensive coverage, and the aggregation of the totality of evidence, research synthesis has emerged as an important and independent research genre or paradigm in applied linguistics. Sin-Wang Chong's edited volume presents a full repertoire of approaches to research synthesis, offering an extensive and in-depth discussion of the challenges, options, and strategies involved in each type of synthesis. The volume is poised to have a profound influence by providing guidance and direction that enable and empower research synthesists to summarize, evaluate, and integrate accumulated knowledge and evidence, thereby advancing the field in ways that would not otherwise be possible."

Shaofeng Li, *The Hong Kong Polytechnic University*

"This volume convincingly demonstrates how research synthesis has gained considerable momentum through diverse methods and applications, offering an impressive variety of in-depth perspectives based on the contributors' roles and expertise. I highly recommend it to anyone interested in how systematic approaches to reviewing the empirical literature, coupled with ethical and multilingual stances, facilitate efforts to evaluate theory, illuminate methodology, and inform practice across the field of applied linguistics. In this era of artificial intelligence, this book serves as a timely reminder of the need to reclaim the knowledge lost in information through collaborative human effort."

Daniel O. Jackson, *Kanda University of International Studies*

ISBN 978 90 272 3769 9



John Benjamins Publishing Company