Hamish Chalmers*, Jess Brown and Anastasia Koryakina

Topics, publication patterns, and reporting quality in systematic reviews in language education. Lessons from the international database of education systematic reviews (IDESR)

https://doi.org/10.1515/applirev-2022-0190
Received December 6, 2022; accepted December 28, 2022; published online January 13, 2023

Abstract: The International Database of Education Systematic Reviews (IDESR.org) contains summary records of published systematic reviews in education and protocols for unpublished reviews and reviews in preparation. During its pilot phase, IDESR is concentrating exclusively on curating systematic reviews in language education. IDESR makes ready access to extant evidence syntheses for researchers, who can use this information to assess the strength of the warrant for any proposed new primary research and/or additional evidence syntheses. By using IDESR to publish review protocols prospectively, review authors commit to high standards of transparency and rigour in producing their research. We have used the data held in IDESR to assess the topics, publication patterns, and reporting quality in the language education literature. We found (i) that language education has seen exponential growth in systematic reviews of research; (ii) that a variety of topics have been addressed, but those related to educational technology have dominated; (iii) that reviews are published in a wide range of outlets, going beyond language education journals; and (iv) that there is room for improvement in the quality of reporting evidence syntheses in language education.

Keywords: language education; reporting quality; systematic reviews

***Corresponding author: Hamish Chalmers**, Department of Education, University of Oxford, 15 Norham Gardens, Oxford, OX2 6PY, UK, E-mail: hamish.chalmers@education.ox.ac.uk. https://orcid.org/0000-0003-2687-9722
Jess Brown, School of Sociology, Politics and International Studies, University of Bristol, 3 Priory Road, Bristol, BS8 1TZ, UK, E-mail: jess.brown@bristol.ac.uk. https://orcid.org/0000-0001-8953-6626
Anastasia Koryakina, EPPI Centre, Social Science Research Unit, UCL Social Research Institute, 10 Woburn Square, London, WC1H 0NS, UK, E-mail: a.koryankina@ucl.ac.uk. https://orcid.org/0000-0003-3274-6440

1 Introduction to IDESR

To understand the state of our knowledge on any topic of research it is necessary to consider the findings of multiple studies together. The most robust way to do this is to conduct a systematic review. The best systematic reviews use transparent and replicable methods to search for relevant evidence; screen reports for eligibility against clear inclusion and exclusion criteria; extract data and appraise study quality; and synthesise the findings of all the research that has addressed the same or similar questions. Good quality systematic reviews help researchers and practitioners to understand what is already known about a topic, and thus inform practice and steer decisions about new avenues for research.

1.1 A database dedicated to systematic reviews in (language) education

The importance of referring to published systematic reviews before conducting new primary research is becoming more widely understood in the field of language education. As we will demonstrate, there has been exponential growth in published systematic reviews in the field, output approximately doubling every five years since 2005. Researchers and practitioners are increasingly drawing on systematic reviews to shape practice and inform research agendas. However, locating relevant systematic reviews can be time consuming and difficult. Organisations that prioritise systematic reviews in the social and educational sciences —the EPPI-Centre at the UK's Institute of Education (eppi.ioe.ac.uk), the American Institute of Education Sciences' What Works Clearinghouse (ies.ed.gov/ncee/wwc/FWW), and the international Campbell Collaboration (campbellcollaboration.org), for example— are obvious places to start when looking for this type of research. However, systematic reviews are published in a host of different outlets.

To be confident that they have located all relevant reviews in their field of interest, scholars face a daunting task if they must locate and search all relevant databases, websites, and journal archives. Because we believe that straightforward access to information about published systematic reviews is important, and because we consider expending large amounts of time searching for and locating relevant research is an avoidable waste of researchers' time and effort, we created IDESR, the International Database of Education Systematic Reviews (IDESR.org).

IDESR is a free-at-the-point-of-access electronic database of published systematic reviews in education, and a clearinghouse for prospective publication of the protocols for planned and ongoing systematic reviews. IDESR is currently in

a pilot phase, concentrating solely on systematic reviews in language education, but as the platform develops will include systematic reviews in all fields of education. IDESR contains the bibliographic information (full reference, title, authors, keywords, and abstract), and a link to original reports of eligible systematic reviews. We describe the processes we used to define eligible literature and populate IDESR in Section 2 of this paper, before presenting an analysis of the contents of the database in Sections 3 and 4.

1.2 IDESR and prospective registration of systematic review protocols

In addition to providing a database dedicated to systematic reviews, IDESR provides a mechanism by which review authors can prospectively publish protocols for their reviews.

1.2.1 What is prospective protocol registration, and why is it important?

As with any research study, a detailed protocol that describes in advance the process and methods that will be applied helps to ensure that all team members involved in producing the research have a shared understanding of the methods and processes to be used (Newman and Gough 2020; Petticrew and Roberts 2006). In the case of systematic reviews, protocols provide a rationale for the review, a specification of the questions that it intends to address, and the methods by which evidence will be located, appraised, and synthesised. When a public record of an ongoing systematic review is available, reviewers planning to conduct a new review can see if a similar review is already in progress. If so, they can redirect their efforts elsewhere, rather than unnecessarily duplicating work. Making the protocol more widely available also means that, should a review need to be replicated or updated, there is a record of the methods and processes by which to do so faithfully (Tai et al. 2020).

Prospectively publishing protocols also helps to protect against biases creeping into the review process. Publication bias (where reviews with unflattering or unexciting results are less likely to be submitted and accepted for publication), for example, can be forestalled by ensuring that once a review is underway it is not lost from view (Petticrew and Roberts 2006). Outcome switching, selective outcome reporting, and modifications to eligibility criteria mid-way through a review are also sources of bias. For example, if during the course of the review process the trajectories of pre-specified outcomes depart from a 'desired' result, reviewers may choose not to report those outcomes or to switch to other outcomes that are a better

fit with their pre-existing assumptions about a topic. This is scientific misconduct (Chalmers 1990), and has been evidenced in both primary and secondary research in applied linguistics (de Bruin et al. 2014; de Bruin and Della Sala 2019; Isbell et al. 2022; Lindstromberg 2022). Prospective registration can help to guard against misconduct of this sort by providing a permanent, publicly available record of what was planned, which can then be compared against what was reported. Where discrepancies exist, journal editors, peer reviewers and readers can assess whether these were likely to have introduced bias.

1.2.2 Registering a protocol on IDESR

IDESR uses Preferred Reporting Items for Systematic Reviews and Meta-analysis Protocols (PRISMA-P) (Shamseer et al. 2015) as the template for the protocols that it publishes. The publication mechanism is based on PROSPERO (crd.york.ac.uk/prospero/), a protocol registry for systematic reviews in healthcare. PRISMA-P is a checklist of items that should be included in any systematic review protocol. The IDESR version of this checklist asks reviewers to state:

- The title.
- Review question.
- Rationale.
- Inclusion criteria.
- Information sources.
- Search strategy.
- Data management process.
- Selection process.
- Data collection process.
- Data items.
- How risk of bias/trustworthiness will be judged.
- How data will be synthesised.
- How confidence in cumulative evidence will be assessed.
- Sources of funding, and
- The role(s) of funders in the production of the review.

Users of the site create an IDESR account, then enter the relevant information about their planned review into an online version of the checklist. The protocol is then submitted for review. The reviewer checks the document for completeness, then either refers the protocol back to the authors for revision or publishes the protocol in IDESR. On publication, protocols are assigned a unique ID and permanent URL, which review authors should include in the report of their review.

To maintain the integrity of the process, IDESR requires that protocols are registered before work on the review has progressed beyond the search phase. If authors deem it necessary to modify their protocol they can do so, stating, and justifying, the reasons for the modification. The updated protocol is date-stamped and republished, providing an audit trail to keep authors' methodological choices transparent. On completion of the review, authors are expected to update the record with a link to the published report, which is then added to the main IDESR database.

Prior to the creation of IDESR, education lacked a dedicated space for systematic reviews and their protocols. While other similar resources address some of the same objectives as IDESR, such as Plonksy's (n.d.) online *Bibliography of Research Synthesis and Meta-analysis in Applied Linguistics* and Open Science Framework (osf.io), IDESR differs from them by its exclusive focus on education. Plonksy's resource, for example, takes a broad view of synthesis in applied linguistics to include reports of studies relating to non-educational outcomes. OSF serves as a repository for all manner of different types of document and data in all fields of research. IDESR aims to streamline and unclutter the process of protocol registration and retrieval for educational researchers. It is our hope that scholars in education will find IDESR to be a well-focused resource that fosters rigorous, open, and transparent scientific conduct in this field.

In the remainder of this paper, we describe the process of curating the contents of IDESR, and present an analysis of those contents.

2 Creating IDESR

2.1 Terms of reference

In the pilot phase of creating IDESR, our aim was to locate, assess for inclusion, and record the bibliographic information of every available systematic review in language education. Systematic reviewing in language education, albeit a growth area, is not as mature as it is in some fields. Some reviews adopt detailed and rigorous methods, while others can be much less meticulous in both conduct and reporting. Given the variation in how evidence syntheses are conducted in the field, and wishing to be as inclusive as reasonable at this stage in the development of IDESR, we interpreted the term 'systematic review' generously. We considered eligible for inclusion topic-relevant syntheses that included a methods section. Moreover, we considered 'systematic review' to be a superordinate term, subsuming different types of review designed to suit different purposes. For example, we considered

scoping reviews, rapid reviews, and narrative reviews, and so on (see Grant and Booth (2009) for a typology), to all constitute systematic reviews.

We expect that as the field matures IDESR will become more discriminatory in what it accepts as eligible for inclusion; prioritising well designed systematic reviews over less rigorous approaches. Indeed, it is hoped that increased interest in evidence synthesis in our field, alongside increased use of IDESR's protocol registry, will encourage production of better designed and more completely reported reviews. We describe in more detail the reporting quality of the reviews included in IDESR in Section 3, and identify areas where our field can improve.

Having established the methodological criteria, we operationalised 'language education' to mean a review with at least one second/additional language learning- or teaching-related outcome.

2.2 Familiarisation with the field

We have noted that systematic reviews in language education vary in the way they are conducted, and we have acknowledged variation in the way that they are named. Prior to conducting our search, we were also aware that some publications likely to meet our definition of a systematic review do not use any of those terms to describe themselves. Most notably is 'meta-analysis', which is a statistical technique that can be used to synthesise evidence derived from systematic methods of selecting literature, but which says nothing *per se* about those methods. Nonetheless, there are publications that adopt some or all of the methodological approaches to systematic reviewing, but which call themselves only 'meta-analyses'. In addition, we were aware of potentially eligible publications describing themselves as state-of-the-art-reviews, evidence syntheses, and so on. We thus took an iterative and cautious approach to searching for literature.

We started by familiarising ourselves with common terminology, to generate search terms that would reflect norms and variations in the field. We did this first by hand searching every issue of the journal *Language Learning*, from its inception in 1948. Every article was assessed for eligibility against our two inclusion criteria. As well as locating eligible publications, this helped us to understand the way in which systematic reviews are discussed.

We repeated this process for a selection of other key journals (see Table 1). However, since our hand search of *Language Learning* located very few systematic reviews in the earlier years of publication, we restricted the hand search to issues published in or after 2015. This iteration of the hand search revealed additional eligible publications and provided us with more information about commonly used terms.

Table 1: Journals and databases searched.

Journals	Databases
Language teaching[a]	ERIC education collection (incl. ERIC) (ProQuest)
The modern language journal[a]	JSTOR
Language learning	Linguistics and language behavior abstracts linguistics collection (ProQuest XML)
Language learning journal	ProQuest dissertations & theses global
Applied Linguistics[a]	
Applied psycholinguistics[a]	
Annual review of applied linguistics	
International review of applied linguistics in language teaching[a]	
International journal of applied Linguistics[a]	
Studies in second language Acquisition[a]	
Canadian modern language review[a]	
TESOL journal[a]	

[a]hand-searched journals (electronic searches were also re-run on these journals).

2.3 Electronic search strategy

Following the familiarisation activity, we created a list of search terms based on our findings (Table 1). We selected key journals and databases (Table 2) and ran searches using those terms. The titles and abstracts of retrieved records were screened for eligibility, then original reports of promising records were located and reviewed in full. We extracted the bibliographic information, the abstract, keywords, and the link to the original report, of those that met our criteria and entered them into the IDESR database.

2.4 Other methods of locating potentially eligible reports

In addition to searching the published literature, we wrote to colleagues in education and applied linguistics, and posted requests on the discussion forums of relevant

Table 2: Search terms[a].

(Language OR language learning OR MFL OR ELL OR EEL OR ELT OR TESOL OR TESL OR FLL OR FL OR FLES OR L2 OR L+ OR L3 OR CLIL OR EMI OR CBI OR immersion OR CBLT OR DLE OR DLI OR BE OR bilingual OR multilingual) AND (systematic review OR evidence map OR gap map OR scoping review OR state of the art review OR meta-analysis OR synthesis OR bibliographic review)

[a]Where databases and journal archives allowed, Boolean phrasing was applied. Where it was not, search terms were entered manually in turn.

professional organisations. We also hand searched the websites of the Campbell Collaboration, the EPPI-Centre, the Education Endowment Foundation, and the What Works Clearinghouse.

When IDESR was launched, we included a mechanism for people to alert us to publications we had missed. Using the 'Tell us about a review we don't have' button displayed on the main pages of the website, IDESR users can send us information about potentially eligible reports.

Finally, the protocol registry in IDESR is designed to be a virtuous circle. When the completed review is published, authors are asked to update the record with its bibliographic information, which is then added to the main database. If, as we hope, evidence synthesists come routinely to use IDESR to register their protocols, this feature will help to ensure that the database maintains an up-to-date record of related research.

2.5 Limitations

While we tried to be meticulous and thorough, and designed in redundancies to minimise oversights, we consider the curation of the database to be a work in progress. Time and resourcing constraints mean that there will be systematic reviews that we have missed. As the project develops we will review our search methods and backfill anything that has escaped our attention. In addition to expanding the search strategy to other databases that may contain relevant records, we intend to search sources of grey literature, such as university research repositories, and government and NGO websites. In time, and through international partnerships, we also intend to include non-English-language reports.

Despite these limitations, we feel that we have curated a sufficiently representative body of literature for the period covered to provide useful and meaningful information for the analyses reported in the next section of this report.

3 Analysing the contents of IDESR

Curating the database has generated a valuable description of the characteristics of systematic reviews in language education. We have begun to explore the possibilities that this dataset holds for understanding this field of research; what the topics of relative priority and relative neglect are; where systematic reviews are being published; and the extent to which they adhere to best reporting conventions.

3.1 Aims and research questions

In Section 1 we described the importance of referring to the current state of our knowledge when planning new systematic reviews and new primary research. In addition to helping to inform researchers' and practitioners' understanding of the world and how this understanding can be operationalised, having an accurate picture of the extent of our existing knowledge helps us to decide if new research on the same topic is warranted. When researchers have an accurate understanding of what is already known, they are in a position to decide whether additional research will, for example, help to consolidate that understanding through replication, address shortcomings or 'gaps' in the current evidence, or, alternatively, constitute a misuse of resources that could be more helpfully directed elsewhere (see Issacs and Chalmers submitted for publication). The first port of call for researchers planning new research, therefore, should be to consult relevant systematic reviews, assess them for quality and completeness, and use that understanding to inform their research agendas.

In this section we aim to assist that process by presenting analyses of the characteristics of the systematic reviews held on IDESR. Specifically, we aimed to address the following questions:

1. What are the patterns of publication of systematic reviews in the field of language education, i.e. How has the volume of published systematic reviews changed over time? Where have systematic reviews been published? And, what topics have been the subject of synthesis?
2. To what extent do systematic reviews in language education adhere to best reporting conventions?

3.2 Method

3.2.1 Data items

Once the database of systematic reviews had been compiled, we extracted the following data: the name of the journal or other mode of publication; publication date; and the author-supplied keywords. This provided us with an overview of the basic characteristics of reviews in the field. To assess the extent to which authors have adopted best reporting conventions in this body of literature, we audited every review for its adherence to the PRISMA 2009 reporting guidelines (Moher et al. 2009).

3.2.2 PRISMA reporting guidelines

PRISMA is a checklist of items that should appear in any systematic review report. While originally created by review authors, methodologists, practitioners, journal editors and consumers in healthcare, the importance of the guideline has been widely acknowledged by evidence synthesists in other fields. It is recommended as a framework for reporting systematic reviews in the social sciences (Gough et al. 2012), in education (Zawacki-Richter et al. 2020), in doctoral and master's research (Boland et al. 2014) and in applied linguistics (Csizéret al. 2022). It is, thus, a widely recognised and well-respected tool. The full checklist, with elaboration, can be downloaded from https://prisma-statement.org//documents/PRISMA%202009%20checklist.pdf. The items are summarised in Table 3.

Table 3: Items on the PRISMA 2009 checklist (Moher et al. 2009).

Section	Item #	Item name
Title and abstract	1	Title
	2	Structured summary
Background	3	Rationale
	4	Objectives
Methods	5	Protocol registration
	6	Eligibility criteria
	7	Information sources
	8	Search
	9	Study selection
	10	Data collection process
	11	Data items
	12	Risk of bias
	13	Summary measures
	14	Synthesis of results
	15	Risk of bias across studies
	16	Additional analysis
Results	17	Study selection
	18	Study characteristics
	19	Risk of bias within studies
	20	Results of individual studies
	21	Synthesis of results
	22	Risk of bias across studies
	23	Additional analysis
Discussion	24	Summary of evidence
	25	Limitations
	26	Conclusions
Administrative	27	Funding

We focus on reporting quality in the analyses that follows, but we consider that reporting quality may also be a useful proxy for methodological quality. The absence of a PRISMA item in a report does not necessarily mean that that item was absent from the method when the review was prepared (Pussegoda et al. 2017). Nonetheless, maintaining the strength of the perception that systematic reviews provide the best evidence to inform decision making (Stewart et al. 2012) relies on authors maintaining the highest standards when they prepare and report them. We feel it is justifiable to conclude that omissions in reporting items widely agreed to be essential in the systematic review process might be interpreted to mean that those items did not feature in the methods (Menon et al. 2022). The authors of PRISMA agree, using the importance of assessing the risk of bias of eligible studies when preparing a systematic review as their example: "… the failure of a systematic review to report the assessment of the risk of bias in included studies may be seen as a marker of poor conduct, given the importance of this activity in the systematic review process" (Moher et al. 2009:2). The same point stands for other items on the checklist. Adherence to PRISMA, therefore, can be considered an indicator of methodological quality as well as a direct measure of reporting quality.

The choice to use PRISMA 2009, rather than its more recent update, PRISMA 2020 (Page et al. 2021) was simply because curation of the data used in our analyses was carried out in 2020, and the analysis of its contents shortly afterwards. That is, before PRISMA 2020 had been published.

Despite systematic reviews being a feature of research in language education for at least a quarter of a century, with a notable uptick in the period immediately following the first iteration of PRISMA in the 2010s (see Section 4.1, below), use of the guideline does not appear to be mandated by journals, and it is rare to find reference to PRISMA in reports of reviews. While it seems likely that anyone with a specific methodological interest in systematic reviews must be aware of PRISMA, it appears that PRISMA is not yet imbedded in practice. Our analysis, therefore, is not intended to criticise review authors when they do not report items on the checklist. Rather, it is intended to highlight where our field can improve its practices by highlighting areas that have been agreed to be important, but which have not yet become standard in language education.

3.2.3 Data extraction

Data extraction was conducted by the authors and three graduate students at the University of Oxford. All have experience of conducting systematic reviews and expertise in the fields of applied linguistics, education, or both. The body of literature was divided among the team, who read through each report and

recorded whether the relevant information was present or absent for each item on the PRISMA checklist.

To help ensure common understanding of the process, all team members independently applied the checklist to a random sample of 10 reviews. We then compared the results in a training workshop, and inconsistencies and general queries about the process were resolved through discussion. When queries arose during the independent work, the team member and the first author discussed and resolved them.

3.2.4 Data analysis

The data items were coded as 1s and 0s (1 = present, 0 = absent) for statistical analysis. This gave rise to a total score of up to 27 for each review and an overall percentage for each item across the body of literature. These data were used to generate the descriptive statistics that follow.

4 Results

Our search resulted in 307 eligible reports. Three of these were available as pre-prints in 2020 (when we conducted the audit) but were not formally published until 2021. We have included data from these publications in all analyses below, though we have not used the 2021 pre-prints in analysis of publication frequency, as this year is incomplete.

4.1 Publication frequency

Figure 1 illustrates the number of systematic reviews published in five-year increments from 2001 to 2020, and a category for reviews published prior to 2001. There were four reviews published, between 1985 and 2000. The 2000 report was Norris and Ortega's, *Effectiveness of L2 Instruction: A research Synthesis and Quantitative Meta-Analysis* (Norris and Ortega 2000), a paper that is widely regarded as 'seminal' in the development of research syntheses in applied linguistics (Goo et al. 2015; Shin 2010). It is perhaps no coincidence that the period immediately after its publication marks the beginning of serious interest in systematic reviews in language education. Between 2001 and 2005, 19 reports were published, with a further 22 between 2006 and 2010. This included five in Norris and Ortega's (2006) edited volume *Synthesizing Research on Language Learning and Teaching*, which built on their earlier work.

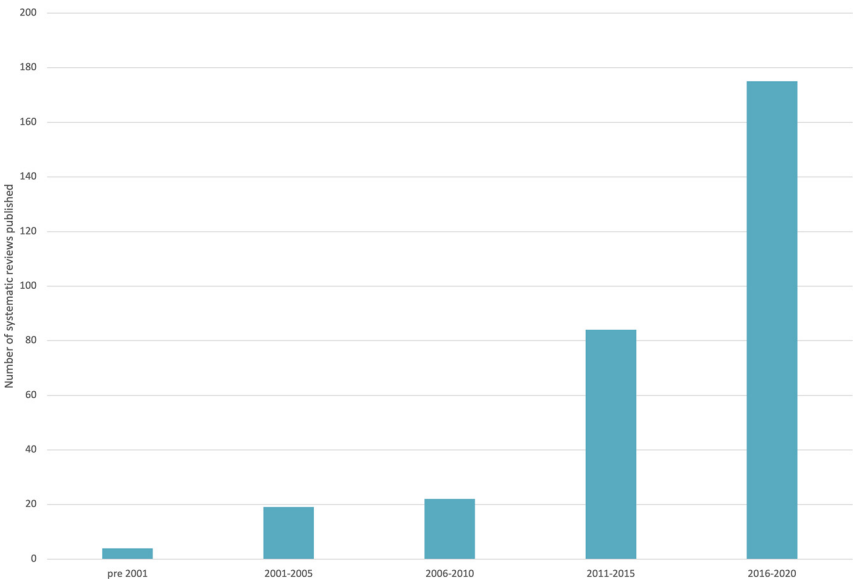

Figure 1: Number of systematic reviews in language education by publication period.

If Norris and Ortega can be credited with kick-starting interest in the field, then the truth of that statement starts to mount from 2011. With 84 reports in the period 2011–2015, the publication rate of systematic reviews was nearly double that of the entire period that preceded it. This doubles again from 2016 to 2020, during which 175 systematic reviews were published, 58% of the entire body of literature.

4.2 Publishers

In total, systematic reviews were published in 136 different outlets. Most reports (n = 270, 88%) were published in scholarly journals, but a small number were published by research centres (n = 7), as government reports (n = 2), as conference proceedings (n = 7), as book chapters (n = 5), and as master's or doctoral theses (n = 16). The majority (n = 107, 79%) have published only one or two reviews. Figure 2 illustrates the frequency of publication by journal, for those that have published three or more reports.

Language Learning leads the field in more ways than one. Having published Norris and Ortega's seminal paper as its first systematic review in 2000, *Language Learning* has built consistently on the momentum it helped to catalyse. It has

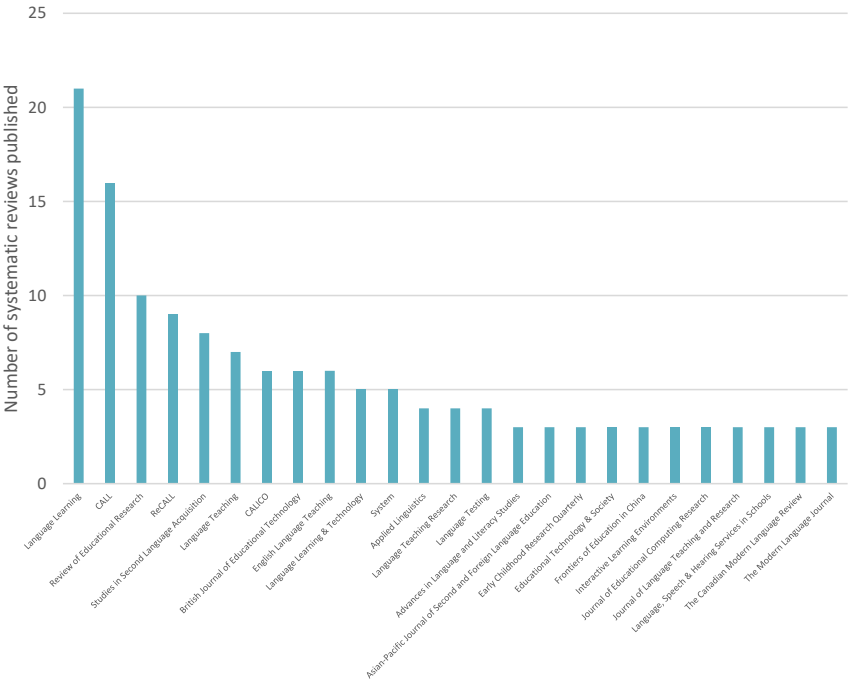

Figure 2: Frequency by publisher.

published 22 reports; three in the period 2001–2005, four in the period 2006–2010, six in the period 2011–2015, and seven in the period 2016–2020 (with one review published formally in 2021). *Computer Assisted Language Learning (CALL)* is the second most frequent publisher, with 16 reports. *Review of Educational Research* published the oldest review in the database (Willig 1985) and two reviews in 2005. Otherwise, most of its output (n = 7, 70%) has been in the most recent decade represented. After *ReCALL* with nine reviews, *Studies in Second Language Acquisition* with eight, *Language Teaching*, *CALICO*, and *British Journal of Educational Technology* each with six (31, or 78%, of which were published since 2010), the frequency for the remaining 16 journals is somewhat similar, with most (n = 11, 69%) having published three reviews.

If *Language Learning* can be said to have fired the starting pistol, the field of educational technology appears to have taken the baton. Of the 25 journals in Figure 2, a third (n = 9) are concerned with language learning and technology. This relative dominance of the field is repeated in the remainder of the literature. In addition to the more prolific language/technology journals, there are many other smaller technology-focussed journals in the dataset. In total, the aims and scope of

31 (23%) of the publishers were explicitly concerned with language learning and technology or educational technology more generally. Collectively these journals have published 84 (28%) of the reviews in the database.

The aims and scope of the remaining publishers range from the expansive (*Open Journal of Social Science*, *AERA Open*) to the niche (*Journal of Special Education Apprenticeship*, *Journal of English for Specific Purposes at Tertiary Level*). A little under half of publishers (n = 63, 46%) were directly related to applied linguistics. Twenty-two (16%) can be described as generalist education publishers. The final 20 (15%) represent a variety of different focuses.

We can see from this analysis that systematic reviews in language learning are valued across the field of education, are not confined to (or even mostly found in) specialist language learning journals, and are of interest to scholars working in a variety of related disciplines. When embarking on new syntheses or new primary research, therefore, language education researchers should be aware that their search for existing systematic reviews to inform their planning should not be confined to the most obvious outlets. When they are considering where to submit completed reviews for publication, they should likewise be open to alternatives.

4.3 Topics

To assess the range of topics subjected to synthesis, we harvested the author-supplied keywords (where available) for each review. In total, 665 different keywords had been supplied. The vast majority (n = 490, 74%) appeared only once in the dataset, 88 (13%) only twice, with the remaining 87 (13%) keywords appearing between three and 85 times. The top three keywords were 'meta-analysis' (n = 85, 13%), 'review' (n = 66, 10%), and 'systematic review' (n = 33, 5%). Together, these represent 61% of the literature. It is unsurprising that, in a body of literature consisting entirely of systematic reviews, these terms are the most frequent. In the analysis, these keywords (along with a small number of others, such as 'effect size', 'empirical research', 'research methods') were removed because they related to methodology rather than topic.

While it might be tempting to interpret the remaining 650 or so keywords as reflecting very broad and varied topics that have been the subject of synthesis in language education, the truth is more mundane. Many of the keywords were very slight variations on each other; 'language education', 'language instruction', 'language pedagogy', and 'language teaching', for example. Therefore, we aggregated keywords that reflected the same or very similar focuses into clusters to illustrate the range of topics covered. For example, we collapsed 'Computer Assisted Language Learning' (CALL), 'Mobile Assisted Language Learning' (MALL), 'Computer

Mediated Communication' (CMC), and other similar terms into a cluster we called 'technology-mediated language learning'. This process resulted in 34 different clusters. The names and frequencies of these clusters is illustrated in Figure 3.

That two of the three largest clusters are 'English Language Teaching' and 'Language Teaching/Learning' is unsurprising given our inclusion criteria. By examining the other clusters in Figure 3, we can gather more illuminating information about what topics have been the subject of synthesis. The dominance of educational technology is reflected by keywords in 'Technology-mediated Language Learning', the largest cluster. There is then considerable drop-off to keywords related to individual differences, formal instruction, assessment and feedback, and bi/multilingualism, each appearing between 61 and 30 times. The smallest circles in the figure show where those larger categories diverge, to provide a better sense of the variety in the literature. Here we see topics such as literacy,

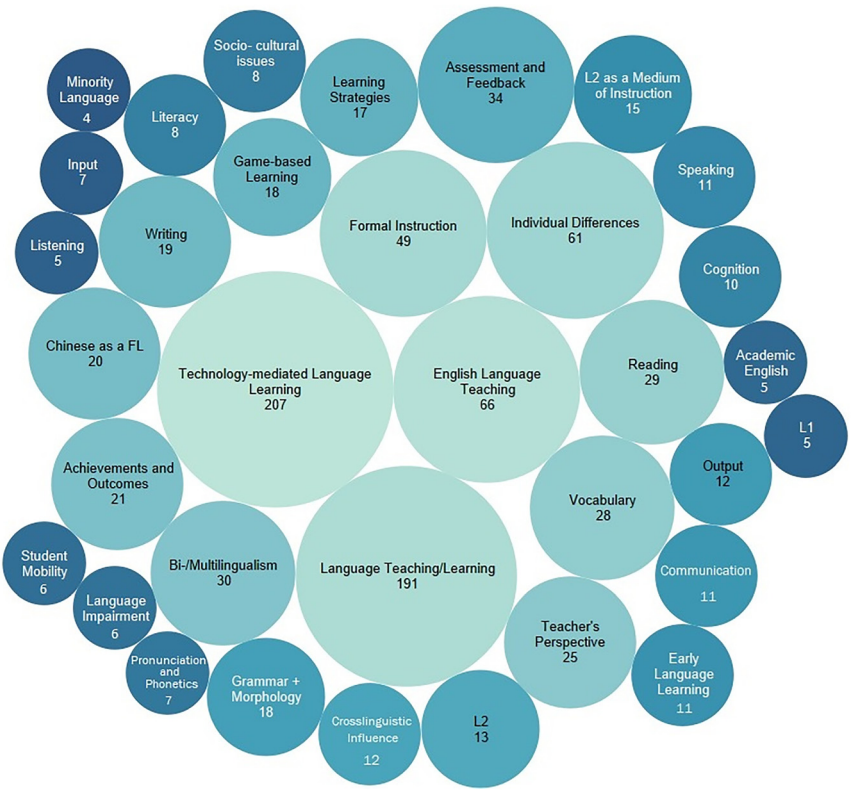

Figure 3: Frequency of topic clusters in the database.

pronunciation and phonetics, academic English, cross-linguistic influence, listening, minority languages, and so on. We will not belabour the analysis by repeating what is self-explanatory in the figure, suffice it to say, there is a wide variety of topics, some that are better represented than others.

These findings will allow prospective systematic reviewers to get a sense of the extent to which their field of interest is already represented in the literature, assess the quality of any relevant reviews, and only then decide whether additional work on the same topic is warranted. If they find that it is not, they can instead make positive contributions to their field by identifying where gaps in the evidence exist and choosing an area of focus based on that assessment.

4.4 Reporting quality

In the final part of our analysis, we present the results of our audit against PRISMA. Figure 4 shows the proportion of reviews in IDESR that report on each item in the checklist against the proportion that do not.

Among the most regularly reported items are those that tend to appear in reports of research of any kind. That is, background and rationale, a statement of objectives, and a conclusion. However, even among these most basic conventions there are instances where reviewers have not addressed them, or addressed them so perfunctorily that they do not meet the standard expected in the field. For example, a relatively small but nonetheless alarming number (19%) did not have an explicit statement of the objectives of the review and/or the questions being addressed, leaving the reader to infer what the reviewers were trying to accomplish. In articulating the background and rationale for their reviews, most authors provided quite detailed reviews of background literature (including, in some cases, the findings of previous relevant systematic reviews). However, some (4%) merely provided a definition of terms and little more. Similarly, a small proportion (6%) did not provide a clear concluding statement about their findings.

About two-thirds (68%) included the terms ‘systematic review’ or ‘meta-analysis’ in their titles. We understand, given our relatively liberal interpretation of what constitutes a systematic review, that some authors may not have realised that their research constituted such and therefore did not use these terms. We hope that increasing understanding of best practice in the discipline will lead authors to routinely include one or both of these terms in the titles of their reviews.

A statement about funding was absent in about two-thirds (68%) of the reviews. The importance of declaring any potentially competing interests is generally well understood by researchers. Whether a review was funded, and if so by whom, helps when assessing the potential for biases to influence the way the review was

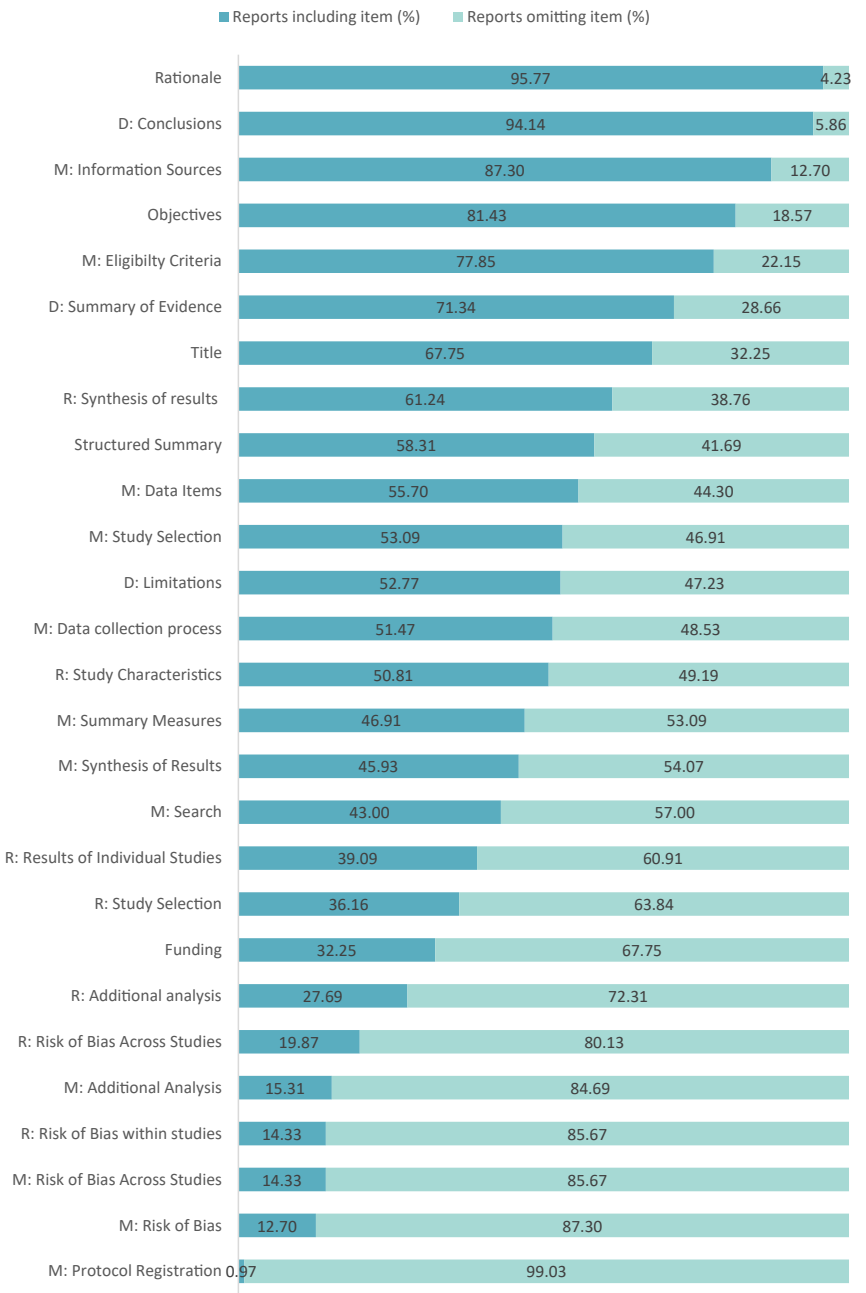

Figure 4: Frequency of reporting methodological characteristics by PRISMA checklist item (n.b. *M* = methods, *R* = results, *D* = discussion).

conducted and reported, even if this means only that a statement that the review was unfunded is provided.

Good systematic reviews stand apart from other types of review because they pay particular attention to the method by which data were collected and analysed. If reviewers do not fully report their methods, the authority of the review might be questioned. If readers do not know, for example, where reviewers looked for reports of research, or what search terms they used, there is no way to assess how thorough their search was. If reviewers do not assess the trustworthiness of the evidence, readers cannot make an informed assessment of the strength of the overall findings, and so on. Failing to fully report the method also makes it impossible to replicate or update a review.

Figure 4 demonstrates that full reporting of methods is far from routine. The items best represented in the literature are 'information sources' and 'eligibility criteria'. Most (87%) indicated where they searched. While some of these included complete lists of the databases consulted, others were unclear. For example, Li (2015:390) states that "The databases searched *included* LLBA, … [etc.]", and Lee et al. (2014:350) said that they "… searched library-housed databases *including* Educational Resources Information Center, … [etc.]" [emphasis added]. It is not clear whether this means 'included, but not limited to', or whether the listed databases were the sum total of information sources consulted. This might seem like a negligible issue of semantics. Nonetheless, clarity matters. The second most frequently reported methods item, 'eligibility criteria', was reported in 78% of reports, leaving readers of nearly a quarter of the reports to infer what constituted eligible studies.

Related to search strategy, fewer than half (43%) included a list of the search terms used to locate eligible evidence. Of those that did, many did not report the Boolean syntax they used, instead just listing the search terms. In some, we found the same lack of clarity we had observed in the way information sources were recorded. Shintani et al. (2013:303), for example, said that "Some of the keywords used in the database search were … ", inviting the question 'Why only some?'. Erling et al. (2016:299) report "… using a variety of key terms associated with the field (e.g. 'language policy', 'language of instruction' and 'medium of instruction')." Again, this invites the reader to question whether other terms were used in addition to these *exempli gratia*. These are small but important points, and we believe that authors (perhaps with the keen eyes of journal editors and peer reviewers) can easily address oversights of this kind.

Other notable omissions in reporting methods were related to process. Just over half reported the process of selecting the studies, and/or the data collection process (53% and 51% respectively). There are also omissions in reporting which data items were collected (56% reported) and the way evidence was synthesised

(46% reported), including reporting how summary measures were calculated (47% reported) and any additional planned analyses (15% reported). We can perhaps be a little more forgiving of these last three items. PRISMA was designed originally for use in healthcare syntheses. It is more common to find systematic reviews in healthcare concentrating on quantitative data and therefore being amenable to statistical synthesis. While statistical syntheses are not uncommon in the reviews in IDESR, we found many that provided a narrative synthesis instead. This might reflect the rise in new forms of systematic review, to include configurative reviews, systematic maps, scoping reviews, and other more recently introduced approaches (Gough et al. 2012). In these cases, summary measures and statistical analysis plans may be less relevant. Nonetheless, explicitly stating the approach to synthesis helps readers to understand the review. It also helps to characterise the nature of the evidence. For example, if a reviewer has chosen not to use meta-analyses, why was this? Was it because the studies did not report quantitative data in enough detail? Was it because studies were so dissimilar as to make statistical synthesis meaningless? Was it because experimental designs are absent from the literature? And so on. These are important pieces of information. They can inform debates about, for example, the use of reporting guidelines for primary studies to ensure readers (and data synthesists) have all the necessary information to interpret research findings. Similarly, knowing that evidence from experiments is rare in a field will inform the development of research agendas that address this, if necessary.

Two of the least well represented items in the literature related to how risk of bias was assessed, both for individual studies and across the review as a whole (13% and 14% reported, respectively). While 'risk of bias' is a term most associated with experimental designs, trustworthiness appraisal is not confined to experiments and is an important part of any systematic review. Understanding the trustworthiness of research is important to understanding what can be reliably concluded from the evidence generated by it. Failure to report this undermines the contribution a review is intended to make.

Finally, on methods, the least well represented item on the checklist was protocol registration. A mere 10 reports (3%) mentioned having a protocol, only three (1%) made these available; one in an appendix published with the report, one on request via email, and one in online supplementary materials. We found no links to prospectively published protocols. This may be explained by the historical lack of a dedicated space to prospectively publish protocols for systematic reviews in language education. It may also be a result of a lack of awareness of the importance of prospective protocol publication. The existence of IDESR addresses the first of these. Increased interest in systematic review methodology in applied

linguistics, and attention to this important component in systematic review training programmes, may start to address the second.

Moving briefly on to the items in the results component of PRISMA, we found a similar pattern to that of the methods section. That is, the extent to which an item was either present or absent in the methods was generally mirrored in the corresponding item in the results. Items that were specific to the results section were typically not very well represented. Clear and complete reporting of the characteristics of each of the studies from which data were extracted was present in only 51% of reports, and the results of individual studies in only 39%.

4.5 Quality over time

To provide an indication of whether quality of reporting has improved as systematic reviewing has become more common in language education, we grouped the literature into the same time periods used in our analyses of the frequency of publication over time (though we included the reports published in 2001 in the last category, as we are comparing percentages rather than absolute numbers) and calculated the proportion of checklist items reported for each period.

As can be seen in Figure 5, reporting quality has remained fairly stable over time, and relatively incomplete. The total number of reports in the earliest periods is small (four in the pre-2001 period and 19 in the period 2001–2005). Interestingly,

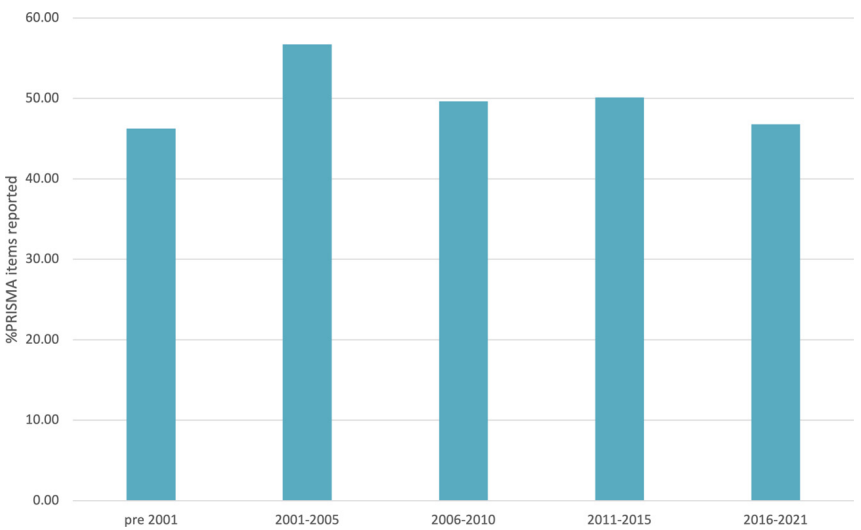

Figure 5: Reporting quality over time.

however, these periods include some of the most completely reported reviews in the database, despite being conducted well before PRISMA was first published. Norris and Ortega (2000) review reported 93% of items, and the EPPI-Centre reviews, all published either in 2004 or 2005, reported between 74% and 85% of items, suggesting that just because a field has developed doesn't mean it has improved.

4.6 Summary of reporting quality

There is wide variation in the extent to which systematic reviews in language education are completely and clearly reported. Overall, the body of literature tends to fall short of best practice, with one or two exceptions. Again, we hope that by highlighting where reporting practices can be improved, we have provided focus for methodological training, and food for thought for review authors and journal editors as practice in the field develops.

5 Conclusions

Systematic reviewing in language education is an area of great activity and growth. The value of increased use of syntheses of this kind promises to improve our assessment of the field, inform policy and practice, and drive new primary research in productive directions. Topics addressed in syntheses are varied, though dominated by those related to educational technology. Systematic reviews are published in a very wide range of outlets, but the reporting quality of this body of research demonstrates that there is room for improvement.

In a 2016 paper entitled *The mass production of redundant, misleading, and conflicting systematic reviews and meta-analyses,* the author rightly describes systematic reviews as "indispensable components in the chain of scientific information" (Ioannidis 2016:486). But he laments that his field (healthcare) has become awash with unnecessary, misleading, and conflicted reviews. He identifies massive, wasteful duplication of effort (for example 185 reviews of antidepressants to treat depression conducted over just seven years); poorly conducted reviews that use fragments of the evidence rather than its totality, inevitably leading to misleading and contradictory findings; and the pernicious influence of vested interests. Even when these poor practices and competing interests are absent, Ioannidis identifies many flaws in the methodological approaches in the literature. He concludes that current systematic reviews "often serve mostly as easily produced publishable units or marketing tools" (485). Few systematic reviews and meta-analyses, he says, are both trustworthy and useful.

If we are to ensure that systematic reviewing in language education does not replicate the situation in healthcare by unnecessarily duplicating work and falling short of expectations for complete and transparent reporting, we must start now. Our recommendations, on the basis of our findings, are that:

- Prospective reviewers should take all efforts to assess how likely their proposed review is to advance our field by referring to, and appraising, what has already been synthesised before embarking on any additional research.
- Plans for new systematic reviews should be documented in carefully prepared protocols, which are then prospectively published.
- Training programmes should embed use of PRISMA as a set of guiding principles for complete reporting of systematic reviews.
- Methodologists should review PRISMA to assess its fit for systematic reviews in language education and develop a field-specific extension if appropriate.
- Until such time, journal editors should expect authors to adopt PRISMA to ensure full and transparent reporting of submissions to their journals.

IDESR can help with these, but the responsibility lies ultimately with authors, journal editors, peer reviewers, and their readerships to insist that the reviews that they produce, publish, appraise, and use conform to the highest reporting standards expected of the discipline.

Acknowledgments: The authors wish to thank Cate Hamilton, Johannes Schulz, Yu Hao, and Julieta Guzman for their help in data extraction, and Gordon Dooley of Metaxis Software Design for building the IDESR database and website.
Research funding: The creation of IDESR was supported by grants from Oxford University Press's John Fell Fund and the Department of Education at the University of Oxford's small grants fund. Funders played no other role in the creation of IDESR nor in the preparation of this report.

References

Boland, Angela, Gemma Cherry & Rumona Dickson (eds.), 2014. *Doing a systematic review, a student's guide*. London: SAGE.

Chalmers, Iain 1990. Underreporting research is scientific misconduct. *JAMA* 263(10). 1405–1408.

Csizér, Kata, Ágnes Albert, & Katalin Piniel. 2022. Editorial: Introduction to the special issue on conducting research syntheses on individual differences in SLA. *Studies in Second Language Learning and Teaching* 12(2). 157–171.

de Bruin, Angela & Sergio Della Sala. 2019. The bilingual advantage debate: Publication biases and the decline effect. In John W. Schwieter (ed.), *The handbook of the neuroscience of multilingualism*, 736–753. Hoboken: Wiley.

de Bruin, Angela, Barbara Treccani & Sergio Della Sala. 2014. Cognitive advantage in bilingualism: An example of publication bias? *Psychological Science* 26(1). 99–107.

Erling, Elizabeth, Lina Adinolfi, Anna Kristina Hultgren, Alison Buckler & Mark Mukorera. 2016. Medium of instruction (MOI) policies in Ghanaian and Indian primary schools: An overview of key issues and recommendations. *Comparative Education* 53(3). 294–310.

Goo, Jaemyung, Gisela Granena, Yucel Yilmaz & Miguel Novella. 2015. Implicit and explicit instruction in L2 learning Norris & Ortega (2000) revisited and updated. In Patrick, Rebuschat (ed.), *Implicit and explicit learning of languages*. Amsterdam/Philadelphia: John Benjamins.

Gough, David, Sandy Oliver & James Thomas. 2012. *An introduction to systematic reviews*. London: SAGE.

Grant, Maria J. & Andrew Booth. 2009. A typology of reviews: An analysis of 14 review types and associated methodologies. *Health Information and Libraries Journal* 26. 91–108.

Ioannidis, John P. A. 2016. The mass production of redundant, misleading, and conflicting systematic reviews and meta-analyses. *The Milbank Quarterly* 94(3). 485–514.

Isaacs, Talia & Hamish Chalmers. In Press. Reducing "avoidable research waste" in applied linguistics research: Insights from healthcare research. *Language Teaching*, Submitted for publication.

Isbell, Daniel R., Dan Brown, Meishan Chen, Deirdre J. Derrick, Romy Ghanem, María Nelly Gutiérrez Arvizu, Erin Schnur, Meixiu Zhang & Luke Plonsky. 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.

Lee, Junkyu, Juhyun Jang & Luke Plonsky. 2015. The effectiveness of Second Language pronunciation instruction: A meta-analysis. *Applied Linguistics* 36(3). 345–366.

Li, Shaofeng. 2015. The associations between language aptitude and second Language grammar acquisition: A meta-analytics review of five decades of research. *Applied Linguistics* 36(3). 385–408.

Lindstromberg, Seth. 2022. P-curving as a safeguard against p-hacking in SLA research. *Studies in Second Language Acquisition* 44(4). 1155–1180.

Menon, Julia L. M., Fréderique Struijs & Paul Whaley. 2022. The methodological rigour of systematic reviews in environmental health. *Critical Reviews in Toxicology* 52(3). 167–187.

Moher, David, Alessandro Liberati, Jennifer Tetzlaff, Douglas G. Altman & The PRISMA Group. 2009. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA Statement. *PLoS Medicine* 6(7). e1000097.

Newman, Mark & David Gough. 2020. Systematic reviews in educational research: Methodology, perspectives and application, In Olaf, Zawacki-Richter, Michael Kerres, Svenja Bedenlier, Melissa Bond & Katja Buntins (eds.), *Systematic reviews in educational research*, 3–22. Weisbaden: Springer.

Norris, John & Lourdes Ortega. 2000. Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis. *Language Learning* 50(3). 417–528.

Norris, John & Lourdes Ortega (eds.), 2006. *Synthesizing research on language learning and teaching*. Amsterdam/Philadelphia: John Benjamins.

Page, Matthew J., Joanne E. McKenzie, Patrick M. Bossuyt, Isabelle Boutron, Tammy C. Hoffmann, Cynthia D. Mulrow, Larissa Shamseer, Jennifer M. Tetzlaff, Elie A. Akl, Sue E. Brennan, Roger Chou, Julie Glanville, Jeremy M. Grimshaw, Asbjørn Hróbjartsson, Manoj M. Lalu, Tianjing Li, Elizabeth W. Loder, Evan Mayo-Wilson, Steve McDonald, Luke A. McGuinness, Lesley A. Stewart, James Thomas, Andrea C. Tricco, Vivian A. Welch, Penny Whiting & David Moher. 2021. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ* 372. n71.

Petticrew, Mark & Helen Roberts. 2006. *Systematic reviews in the social sciences*. Oxford: Blackwell.

Plonksy, Luke. no date. *Bibliography of research synthesis and meta-analysis in applied linguistics*. Available at: https://lukeplonsky.wordpress.com/bibliographies/meta-analysis/ (accessed 7 October 2022).

Pussegoda, Kusala, Lucy Turner, Chantelle Garritty, Alain Mayhew, Becky Skidmore, Adrienne Stevens, Isabelle Boutron, Rafael Sarkis-Onofre, Lise M. Bjerre, Asbjørn Hróbjartsson, Douglas G. Altman & David Moher. 2017. Identifying approaches for assessing methodological and reporting quality of systematic reviews: A descriptive study. *Systematic Reviews* 6(117). 1–12.

Shamseer, Larissa, David Moher, Mike Clarke, Davina Ghersi, Alessandro Liberati, Mark Petticrew, Paul Shekelle, Lesley A. Stewart & The PRISMA-P Group. 2015. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P): Elaboration and explanation. *British Medical Journal* 349. g7647.

Shin, Hye Won. 2010. Another look at Norris and Ortega (2000). *Studies in Applied Linguistics & TESOL* 10(1). 15–38.

Shintani, Natsuko, Shaofeng Li & Rod Ellis. 2013. Comprehension-based versus production-based grammar instruction: A meta-analysis of comparative studies. *Language Learning* 63(2). 296–329.

Stewart, Lesley, David Moher & Paul Shekelle. 2012. Why prospective registration of systematic reviews makes sense. *Systematic Reviews* 1(7). 1–4.

Tai, Joanna, Rola Ajjawi, Margaret Bearman & Paul Wiseman. 2020. Conceptualizations and measures of student engagement: A worked example of systematic review. In Olaf Zawacki-Richter, Michael Kerres, Svenja Bedenlier, Melissa Bond & Katja Buntins (eds.), *Systematic reviews in educational research*, 91–110. Weisbaden: Springer.

Willig, Ann C. 1985. A meta-analysis of selected studies on the effectiveness of bilingual education. *Review of Educational Research* 55(3). 269–318.

Zawacki-Richter, Olaf, Michael Kerres, Svenja Bedenlier, Melissa Bond & Katja Buntins (eds.), 2020. *Systematic reviews in educational research*. Weisbaden: Springer.