Introduction

Meta-analysis is a statistical method for quantitatively summarizing and synthesizing data from multiple studies (Borenstein, Hedges, Higgins, & Rothstein, 2009). It has a growing importance in applied linguistics, as the field develops and matures and the number of studies therein increases accordingly. Although this growth and maturation may be a welcome trend overall, it cannot be assumed that it will lead researchers to enrich their knowledge about language-related matters. One reason is that studies may be produced in a scattershot manner, may become hard to locate, and may not be available when needed (Cooper & Hedges, 2009b). This may make a comprehensive picture hard to achieve, resulting in researchers repeatedly examining topics that have already been investigated before. Another reason is that even if studies are readily available, their findings are often inconsistent, making it difficult to confidently interpret what works and what does not (Lipsey & Wilson, 2001). This could partly be because of the many contextual and moderator variables that come into play across studies; the effects of these variables need to be closely examined and teased out as they influence the relationship between independent and dependent variables.

This chapter discusses the application of meta-analysis in applied linguistics. Meta-analysis can afford broader understanding of the overall findings and characteristics of applied linguistics studies and inform the design of new studies based on the accumulated knowledge in the field. We first describe meta-analysis and compare it to other approaches to summarizing studies. Then we consider how meta-analysis has been used in applied linguistics. Next, we present an example of meta-analysis in applied linguistics while noting issues and challenges. The chapter concludes by discussing the advantages and limitations of meta-analysis and its potential applications to future applied linguistics studies.

Approaches to research synthesis

Cooper and Hedges (2009b) describe research synthesis as the practice of scrutinizing previous studies and relevant theories, combining study findings, examining the scope of generalizability of the integrated findings, explaining inconsistent findings, and offering useful
directions for further research. These steps are essential to draw optimal conclusions about a topic of interest, and can be conducted using several different approaches. Next, we describe a traditional and a more recent approach to research synthesis, based on the discussions in Norris (2015) and Oswald and Plonsky (2010).

**Traditional approaches: narrative review and vote-counting review**

Narrative review and vote-counting review are two well-established approaches to research synthesis. In a narrative review, studies are reviewed based on reviewers’ experience or expertise. In a vote-counting review, in contrast, conclusions are primarily drawn based on the number of studies reporting statistically significant findings. For example, suppose that ten studies are collected comparing the relationship between working memory and second language (L2) reading comprehension. Each study will be examined as to whether the relationship is statistically significant, and the results tallied. If more studies report statistically significant results supporting the effect, one concludes that the finding is sound.

Narrative and vote-counting reviews are common and flexible (Li & Wang, 2018) but limited. One limitation is that studies are collected and reviewed by a human reviewer, generally the author, whose criteria for including/excluding studies may not always be clear (see, e.g., Oswald & Plonsky, 2010). How reviewers reach the conclusions are not often made explicit, either. Another limitation of vote-counting review is that it depends primarily on statistical significance testing, which can change from significant to non-significant according to treatment and sample size even with effects of the same magnitude (see, e.g., Oswald & Plonsky, 2010).

**Recent approaches: systematic qualitative review, systematic quantitative review, and systematic quantitative meta-analytic review**

More recent approaches to research synthesis include not only quantitative but also qualitative methods. Common across these approaches are, first, systematic literature search processes (see Macaro, this volume) and explicitly reports selection criteria for studies into review (Oswald & Plonsky, 2010). Studies are collected using not only the authors’ expertise but also databases, manual searches of literature, the references section of papers, and other researchers’ expertise in the field. Second, this exhaustive search process is bolstered by explicit criteria for inclusion/exclusion of studies, often applied by multiple coders to maintain consistency. Studies screened in this rigorous way are then summarized based on the author’s expertise in qualitative review or numerically in quantitative review.

Quantitative reviews can be divided into meta-analytic approaches and other quantitative approaches. In quantitative meta-analytic reviews, studies are combined regardless of statistical significance, as more emphasis is placed on the aggregated effect size (as opposed to a dichotomous \( p \) value judgement of whether there is an effect or not) and confidence interval (i.e., accuracy or precision of the measurement) across studies. In another type of quantitative review, termed a “quantitative synthetic approach” in In’nami and Koizumi (2014), authors combine collected studies thematically and typically interpret the overall tendency and other features using descriptive statistics such as proportions and frequency. The previously discussed approaches to research synthesis are summarized in Table 20.1. Note that more than one approach can be used in a single study.

Téllez and Waxman (2006) conducted a qualitative meta-synthesis of qualitative studies on effective teaching practices for English language learners in the US (see Finfgeld, 2003). Studies
from five databases were judged in terms of four criteria (e.g., presence of a rationale for choosing participants; sufficient context to support a qualitative study). Of the 50 studies identified, 25 were accepted into their meta-synthesis; these were coded based on theoretical frameworks, and the appropriateness of the frameworks was checked by two L2 specialists. Results yielded several features of effective teaching practices for English-learners in the US; for example, cooperative learning, or “communitarian teaching practices,” was found to be more effective than non-cooperative learning because social interaction is an important aspect of learning language.

Macaro, Curle, Pun, An, and Dearden (2018) synthesized empirical studies in a systematic quantitative and qualitative review of English medium instruction. Studies collected from databases and journals were judged in terms of seven inclusion criteria (e.g., reporting empirical data) and five exclusion criteria (e.g., master’s dissertations) by two experts, narrowing an initial total of 606 studies down to 83. Results showed that research on English medium instruction was predominantly conducted in Asia and Europe (31 of the 83 studies, 33 of the 83 studies, respectively). For more information on systematic review methodology, see Macaro’s chapter in this volume.

In a recent meta-analysis, de Vos, Schriefers, Nivard, and Lemhöfer (2018) quantitatively synthesized studies gathered using databases, journals, and manual reference search to examine the effectiveness of incidental L2 word learning through spoken input as a function of five variables (three substantive and two methodological). Studies were screened using ten criteria, with 30 of 319 studies included. Spoken input was found more effective than no input for incidental word learning using Hedges’s $g$ effect size $= 1.05$, a large effect according to Plonsky and Oswald (2014); 95% confidence interval (CI) $= 0.81$ and 1.28, standard error $= 0.12$, $z = 8.77$, and $p < .001$. Among the five variables examined, learning effect, for example, was positively related to age, with university students outperforming kindergarten and elementary school students. The authors speculated that this was because university students had more experience learning the L2, were more motivated to learn words, and/or had higher levels of intelligence than elementary school students – possible explanations that were confounded with age, thus requiring further research.
Meta-analysis

Meta-analysis is a systematic statistical quantitative approach to integrating a large body of research for meaningful interpretation (Cooper & Hedges, 2009b). As the field of applied linguistics has matured, meta-analytic studies have mushroomed in number. As of this writing (November 2018), Plonsky’s (n.d.) bibliography of research syntheses and meta-analyses in applied linguistics lists 358 references, among which 234 have “meta-analysis” in the title. These numbers were 176 and 142, respectively, as of April 2014, showing the growing popularity of meta-analysis in the field.

Meta-analytic studies in applied linguistics have commonly appeared in Language Learning, which has a “systematic review article” section (for an example, see Plonsky, 2011). Li, Shintani, and Ellis (2013), Ortega (2015), and Plonsky and Oswald (2015) provide accessible introductions to conducting meta-analysis, while the journal Research Synthesis Methods collects the latest methodological developments and Psychological Methods has published on statistical aspects of meta-analysis.

Stages of meta-analysis

Following the six stages of research synthesis by Cooper and Hedges (2009b), we illustrate the use of meta-analysis to summarize studies.

Formulating a problem

The research topic for a meta-analysis should be well researched, with a sufficient number of quantitative studies available (as meta-analysis is a method for quantitative data). Meta-analyzed topics in applied linguistics include working memory (Linck, Osthus, Koeth, & Bunting, 2014), computer-mediated language learning (Lin, 2014), and L2 instruction on reading (Jeon & Yamashita, 2014), vocabulary (Uchihara, Webb, & Yanagisawa, 2019), and grammar (Spada & Tomita, 2010).

Literature search

Once the topic is selected, a systematic literature search through multiple avenues is conducted. Keeping detailed records of decisions made and actions taken at each stage is essential for valid, interpretable results. First, keywords are selected and databases searched (e.g., EBSCO, ERIC, JSTOR, LLBA, ProQuest, PsycINFO; see In’nami & Koizumi, 2010, for database selection). Note that databases may overlap – for instance, ERIC, LLBA, and PsycINFO are independently available but also included in ProQuest. It is important to know which database(s) we are actually searching. Second, reference lists of retrieved studies can be consulted for further related research. Third, conference programs and books can also be manually checked, and authors of relevant studies may be contacted if needed for unreported information on matters such as study design or statistical information.

Evaluating data

Through the procedures discussed earlier, we might obtain an overwhelming number of studies. As Plonsky and Oswald (2012) suggest, over-searching is better than under-searching, but inclusion/exclusion criteria are nevertheless crucial. Inclusion criteria are often based on
study characteristics including publication type (e.g., peer-reviewed journal, book chapter, dissertation, conference presentation), publication year, research domain (e.g., individual differences, pragmatics, pronunciation, and grammar), context (e.g., classroom, laboratory), target language, target population, or study designs (e.g., pretest–posttest, instructional treatment, control/comparison groups), or on statistical characteristics such as means, standard deviations, number of participants, outcome reliability, or effect sizes (Lipsey & Wilson, 2001). All information should be recorded on a detailed coding sheet.

When modifying or developing coding categories, we need to carefully define each category to ensure coder reliability and accurate interpretation of meta-analysis results. For example, a “control group” can be interpreted differently depending on research domain. It can be a group without any contact with or attention to the target structure. Alternatively, it can be a group with only minimal contact with the target structure (e.g., the structure may be used in the participants’ reading assignments but not highlighted). After defining “control group,” we need to consider what to do with a study that does not have a control group. One option is to exclude the study, because we cannot calculate an effect size by comparing a treatment group and a control group. In this case, having a control group becomes an inclusion criterion. A second option is to select a group that received the least instructional treatment as a baseline group to calculate effect sizes for comparison groups. Then, having a control/baseline group becomes an inclusion criterion. The third option is to calculate an effect size by comparing pretest and posttest data. This would make using a pretest–posttest design an inclusion criterion.

Each choice has advantages and disadvantages. For example, when we are specifically interested in differences between a treatment group and a control group, we can confidently exclude studies without a control group, but by losing those studies we ultimately have a smaller number of studies included in the meta-analysis. As for selecting a baseline group, different coders may choose different groups as baselines, affecting coder reliability. Last, comparing pretest and posttest tends to produce larger effect sizes than comparing a treatment group and control group (Plonsky & Oswald, 2012). Thus, the latter two methods (selecting a baseline group and comparing pretest and posttest) may have the benefit of retaining studies but also the risk of decreasing coding reliability. Even if we include pretest–posttest data, we should analyze them separately from treatment group–control group data, since between-group and within-group comparisons are different in nature (Plonsky & Oswald, 2015). After considering such important issues, we should develop clear eligibility criteria with detailed definitions for each category in order to code consistently and interpret the meta-analysis results with confidence. Naturally, researchers should continue revising their definitions while coding further studies, because we often notice some ambiguities in definitions and find the necessity of being more narrowly focused when facing difficulty coding some categories.

Reliability of coding should be checked through double coding, either by a single coder or by multiple coders coding the same studies independently. A single coder needs to code the studies again after a sufficient time period has passed, to minimize memory effects. In either case, we want to make sure that the results of coding are consistent, whether across two occasions for a single coder or across multiple coders. Lipsey and Wilson (2001) recommend that “20 or more studies, with 50+ being more desirable” (p. 86) should be coded to ensure coder reliability. However, in second language acquisition (SLA), meta-analyses with such large numbers of studies are rare; thus, all studies may need to be double coded (Plonsky & Oswald, 2015) after narrowing down based on retrieved abstracts from more than 1,000 studies to a smaller number, such as 30 studies. Coder reliability can be reported as a percentage of agreement or using indexes such as Cohen’s kappa.
Analyzing data

After selecting studies that meet all inclusion criteria, we need to consider what effect size to use for the meta-analysis and then calculate effect sizes for each study. Among numerous types of effect sizes, one characteristic of treatment group–control group comparison is a standardized mean difference, such as Cohen’s $d$ or Hedges’s $g$. If pretest and posttest are compared, Glass’s delta is another option. Such effect sizes can be calculated, for example, based on means, standard deviations, and sample sizes. When relationships between two variables are in focus, correlations such as Pearson’s product-moment correlation coefficient ($r$) are often used. See de Vos et al. (2018), for different calculation of effect sizes depending on study design. Effect sizes can be calculated using dedicated software (e.g., Del Re, 2015; Wilson, 2010, 2018) or comprehensive meta-analysis software (Biostat, 2006–2018; Viechtbauer, 2016).

Important decisions have to be made in calculating effect sizes, such as how to choose effect sizes when there is more than one in a study. One effect size per study is recommended for a meta-analysis to maintain independence of responses (Lipsey & Wilson, 2001). However, many studies in applied linguistics, particularly in instructed SLA, have multiple treatment groups and multiple outcome measures, producing multiple dependent effect sizes. According to Lipsey and Wilson (2001), one option is selecting one effect size among all, whereas another option is averaging all effect sizes into one. However, these methods may wipe out instructional treatment differences, which are the major interest of instructed SLA studies. An alternative option is averaging effect sizes in accordance with the research questions and purposes of the meta-analysis. For example, a meta-analysis that intends to summarize the effect of different types of instruction on L2 learning outcomes in general can average into one all effect sizes from different outcome measures for each instructional treatment. If the goal is instead to summarize the effect of different types of instruction on different L2 learning outcomes, such as spontaneous vs. controlled use of language, two effect sizes can be calculated for each instructional treatment: an average of all effect sizes of the spontaneous outcome measures and an average for controlled outcome measures. Certainly, readers must know how each effect size is calculated to interpret the results of a meta-analysis; therefore, as noted, we need to record and report all decisions made. Polanin, Hennessy, and Tanner-Smith (2017) offer a useful review of meta-analysis packages in R (see Larson-Hall & Mizumoto, this volume) and provide a tutorial on addressing dependent effect sizes (i.e., multiple effect sizes within a single study) using the robumeta package (Fisher, Tipton, & Zhipeng, 2015).

Once effect sizes are calculated for each study, we need to calculate mean effect sizes and other statistics. Effect sizes are sample-size weighted (using the inverse of the sampling error variance) to obtain unbiased effect sizes, because effect sizes from large-sample studies have fewer sampling errors and should contribute more to the final results in the meta-analysis than those from small-sample sizes (see Lipsey & Wilson, 2001, for calculation formulas).

Interpreting results

Before interpreting the results, we may want to confirm that there is little sampling or publication bias in the effect sizes. That is, authors of studies with statistically nonsignificant results are less likely to submit their studies to journals (the file-drawer problem) and effect sizes in meta-analyses may come disproportionately from published studies with significant results. This may yield larger effect sizes and overestimation of summary effects than including unpublished studies would have (e.g., Lipsey & Wilson, 2001). However, unpublished studies may (also) have other issues such
as low research quality or limited circulation. To examine sampling or publication bias, statistical methods such as the fail-safe $N$ or a funnel plot can be used (e.g., Borenstein et al., 2009).

In addition, to ensure that summary effect sizes properly estimate population effect sizes, we can run the homogeneity test, using the $Q$ statistic. A statistically significant $Q$ indicates that summary effect sizes do not estimate single-population effect sizes and that there may be other factors (e.g., study characteristics) affecting effect sizes. In such cases, we can use a random-effects model, a fixed-effect model with moderator variables, or a mixed-effects model (see Lipsey & Wilson, 2001, for further details and calculation formulas). A statistically nonsignificant $Q$, in contrast, indicates homogeneous distribution of effect sizes, which allows us to assume a fixed-effect model, where summary and population effect sizes differ only by sampling error. However, nonsignificant $Q$ does not always guarantee the use of a fixed-effect model because the $Q$-test may fail to reach statistical significance simply due to a small number of effect sizes, based on a small number of studies included, even when there is actually a heterogeneous distribution of effect sizes. More importantly, models should be selected based on theory, not on the results of a $Q$-test (see Borenstein et al., 2009).

When interpreting effect sizes, Cohen (1988) provided useful benchmarks for small ($d = .20$), medium ($d = .50$), and large ($d = .80$) effect sizes, respectively. However, Plonsky and Oswald (2014) suggested alternative benchmarks for SLA studies: $d = .40$ is a small effect, $d = .70$ is a medium effect, and $d = 1.00$ is a large effect for treatment–control comparisons.

There are some caveats when interpreting effect sizes. First, the quality of the original study influences the results of the meta-analysis (the garbage-in, garbage-out problem). If we include studies with methodological flaws, the results may also become weak and unreliable. Second, a large effect size may be due to confounding effects. For example, explicit instruction may produce a large effect size, but close examination may elucidate that the outcome was measured by controlled tests similar to the activities used during the explicit instruction treatment. Such “reactive measures” (Lipsey & Wilson, 2001, p. 158) may not allow us to appropriately interpret effect sizes. Finally, if the number of studies is small, the conclusion may not be strong and the results may need to be interpreted cautiously. See Plonsky and Oswald (2015) for other issues involved in interpreting meta-analytic results.

**Presenting results**

Results should be reported in a comprehensive, easy-to-understand manner, including how the research focus was determined, studies were collected, and results were analyzed and interpreted. Overall results should be presented in relation to moderator variables, and summarized findings across studies should be discussed, followed by identification of gaps to be filled in future research. To effectively communicate results to readers, tables and figures should be used (Plonsky & Oswald, 2015). When important information has to be excluded, it should be posted as supplementary material online (Plonsky & Oswald, 2015).

One common figure type in meta-analysis is a forest plot. Figure 20.1 displays an overall tendency and variations across studies by showing effect sizes for each study (black boxes) and synthesized effect sizes (diamond) with 95% CIs. Figure 20.2 presents a funnel plot, displaying relationships with effect sizes and sample sizes of each study. If publication bias does not exist, effect sizes are symmetrically distributed around the mean effect size.

Figure 20.1 Forest plot of effect sizes of changes in L2 proficiency of Japanese university students learning English before (Group A) and after (Group B) studying abroad, from Hirai (2018). Positive values of Hedges’s $g$ show improvement in L2 proficiency. This figure was created using Comprehensive Meta-Analysis (Version 2; Biostat, 2006–2018).
Figure 20.1  Forest plot of effect sizes of changes in L2 proficiency of Japanese university students learning English
Strengths and weaknesses of meta-analysis

Although meta-analysis is a popular, useful method in applied linguistics, it is not always the best procedure for synthesizing studies. The strengths and weakness of meta-analysis have been covered elsewhere (e.g., see Borenstein et al., 2009; Cooper & Hedges, 2009a; Ellis, 2018; In’nami & Koizumi, 2014; Lipsey & Wilson, 2001; Norris & Ortega, 2006; Oswald & Plonsky, 2010); here, we focus on certain strengths and weaknesses that we believe most merit attention.

One advantage of meta-analysis is that it structures the steps or phases involved in summarizing studies comprehensively and reporting them explicitly and in detail, so that readers can judge the veridicality of the meta-analysis and replicate it themselves (e.g., Lipsey & Wilson, 2001). There are six phases, as discussed earlier; while most of them are common to all synthetic review approaches, meta-analysis has a distinct data analysis phase, with standard procedures for computing effect sizes and integrating them. Conducting a meta-analysis is often a time- and labor-intensive process, and this is particularly true when a large number of studies are available and they vary in research design, statistics reported, and/or contextual variables reported, among other aspects. Structuring these processes may help readers better understand how meta-analytic studies are conducted, without taking them at face value.

A second advantage is that findings from each study are more clearly and comparably presented in meta-analysis (e.g., Norris & Ortega, 2006). More specifically, the degree/magnitude of the difference or relationship between groups, such as the effectiveness of a treatment, is represented by effect size weighted by sample size, and its accuracy or precision is represented by the confidence interval of the effect size. Interpreting this degree of difference and accuracy is more illuminating than just interpreting whether the treatment is effective or interpreting studies with no report on accuracy of what is presented, as is often the case in vote-counting review. Further, when effect sizes calculated for each study are combined across studies, the result is a summarized effect size that is more informative than interpreting each individual study’s effect size. When effect sizes are combined across studies according to variables of
theoretical or methodological interest, the results should indicate how those variables are related to the effectiveness of the treatment.

Meta-analysis has disadvantages or considerations as well, however. One is whether it is sensible to combine studies (e.g., Norris & Ortega, 2006). This is called the “apples-and-oranges problem,” and requires researchers to consider whether the summarized effect size is meaningful. If one is interested in an overall effect of strategy instruction across skills (e.g., in Plonsky, 2011, reading, speaking, and vocabulary), one should combine all studies. If, on the other hand, one’s interest is in a particular skill – say reading – studies on reading strategy instruction should be treated separately and not be combined with those on speaking or vocabulary strategy instruction. Researchers should endeavor to make theoretically – and not necessarily statistically – sensible decisions when considering whether to combine studies.

A second disadvantage is that the quality of meta-analysis depends on the quality of the primary studies (e.g., Oswald & Plonsky, 2010). If the primary studies are well conceived, designed, executed, and reported, they will be invaluable input for the meta-analysis. However, not all studies satisfy all these conditions. For example, studies may be well conducted, but the results may not report the statistics necessary to calculate effect sizes. Further, such information may not be available from the author of the study, and the data may not have been archived. This means that these otherwise potentially important studies instead need to be dropped and/or the missing data imputed in the meta-analysis. If such studies abound and their inclusion, exclusion, or imputation produces different meta-analytic results (sensitivity analysis; see, e.g., Borenstein et al., 2009), one should be cautious when interpreting them.

Finally, although research synthesis processes in meta-analytic studies are structured and reported in detail, which makes those studies more objective and transparent than other review methods, this does not necessarily indicate that they are also more trustworthy than other review methods. One reason is that meta-analysts must make many decisions while conducting their meta-analysis, affecting the results. For example, in the database search stage, researchers must constantly reevaluate what databases are used, what keywords are used, how they are used in combination with wildcards, whether search duration is restricted to a particular time period, whether abstract-only or a full-text search is conducted, and whether studies published in languages other than English are included. In our experience, these decisions surrounding the use of databases are often made by a single author, in contrast with coding effect sizes, which is usually conducted by two or more authors.

Comparisons among three specific studies (Shintani, 2015; Lee, Jang, and Plonsky, 2015; and Li, 2015) can illustrate differences in decisions that meta-analysts can make. Although these studies all appeared in the same special issue on the possibility of the complementarity of narrative review and meta-analysis, and were considered exemplary, the decisions the meta-analysts made were not always the same. For example, only published studies were included in Shintani (2015), whereas both published and unpublished studies were included in Lee et al. (2015) and Li (2015). Publication bias was not checked in Shintani (2015), whereas it was checked but not addressed in Lee et al. (2015) and was examined and found not to influence the results in Li (2015). Inter-coder reliability was 98.2% in Shintani (2015), whereas it was not reported in Lee et al. (2015) or Li (2015).

Another example is the languages in which studies are written. In general, in meta-analyses, articles in English are searched and included. Macaro et al. (2018) also had Chinese speakers on their team, and they conducted keyword searches on abstracts written in Chinese. They reported that they found a large body of studies, although they did not include them in their synthesis. We believe their attempt was worthwhile, because there might always be relevant
studies out there reported in languages other than English – particularly in Chinese, given the growing number of studies published by researchers affiliated with universities in China. As an attempt in this vein, two of the authors of this chapter (In’nami and Koizumi) are currently conducting a meta-analysis on some cognitive variables and L2 reading comprehension. With the help of a Chinese graduate student in China using the China National Knowledge Infrastructure (CNKI: www.cnki.net) – a “Google Scholar of China” – we found 99 studies written in Chinese and published or reported in China. This stresses the importance of having members or assistants on a research team who are familiar with languages in addition to English and of searching for non-English studies.

Thus, the various decisions involved in conducting a meta-analysis will lead to different results and interpretations and suggest the complexity and challenges meta-analysts face.

Implications

As data-analytic choices overall “can be highly contingent on justifiable, but subjective, analytic decisions” (Silberzahn et al., 2018), this will also be true for meta-analysis, where researchers must make decisions on many analytic issues such as types of effect sizes (d, r) and models (fixed effect, mixed-effects, random-effects), as well as treatment of dependent effect sizes. No golden rule exists regarding analytic choices, but there are guidelines and recommendations that one should follow, including APA Publications and Communications Board Working Group on Journal Article Reporting Standards (2008) and Larson-Hall and Plonsky (2015). As these guidelines and recommendations are continually being revised as the field develops, perhaps the best one can do is to report one’s (analytic) decisions explicitly and make the data publicly available so that they can be better analyzed if needed and the results better contribute to the accumulation of knowledge in the field.

Perhaps such a process could be expedited by creating a website where data from published meta-analyses are archived and authors of (newly published) relevant studies can add their own data. These data could then also be edited or revised as more variables are coded, coding mistakes are corrected, and/or more accurate formulas for estimating or combining effect sizes are developed. These updates could be led and monitored by researchers. Such a website could be created using Shiny – an open-source package for R that allows one to create an interactive website. In this way, knowledge in the field can be made more widely available and continuously updated. Meta-analysis is a powerful tool to move applied linguistics researchers in that direction.

References


