CHAPTER THIRTEEN

Research Synthesis

Lourdes Ortega

Research synthesis refers to a continuum of techniques and research procedures that have been developed by social scientists with the aim of reviewing past literature systematically. Simply put, the methodology produces contemporary literature reviews that differ from traditional literature reviews in taking an empirical perspective on the task of reviewing. Syntheses (also known as *systematic reviews* in some fields) investigate and evaluate past findings in a systematic fashion, always explicating the methodology followed in the review so as to enable replication by other reviewers. The approach began in the late 1970s in the fields of psychology, education and medicine and its use has been widespread across the social sciences since the late 1980s. In applied linguistics, the methodology was introduced in the mid 1990s and has witnessed a rapid development particularly since the late 2000s. Table 13.1 lists forty-five article-length syntheses published in the first twenty years of application of this method in our field.

Meta-analysis is probably the best-known technique for systematically synthesizing quantitative research. However, meta-analyses are a restricted form of synthesis that cannot always be conducted. For one, synthesists can engage in meta-analysis only when the body of research to be synthesized is clearly quantitative, that is, experimental, quasi-experimental or correlational. In addition, the methodology of meta-analysis is driven by questions about the magnitude of an effect or causal relationship and involves mathematical ways of summarizing past findings that demand the availability of a large number of studies.

Table 13.1 Article-length systematic research syntheses published in applied linguistics (1994 to February 2014)

Study	Key research question
Thomas (1994)*	How have SLA researchers measured proficiency for research purposes in their studies?
Ross (1998)**	How well does L2 self-assessment work?
Norris and Ortega (2000)**	How effective is L2 grammar instruction, and does type of instruction matter?
Goldschneider and DeKeyser (2001)**	How much of the L2 English morpheme accuracy order can be attributed to frequency and salience?
Masgoret and Gardner (2003)**	How well does amount of motivation predict L2 achievement in Garner's Attitude/Motivation Test Battery?
Ortega (2003)*	What is the relationship between syntactic complexity and L2 writing proficiency?
Roessingh (2004)**	What elements in the design and implementation of ESL programs support successful outcomes for ESL learners?
Dinsmore (2006)**	How different are native and non-native speaking performances in Universal Grammar studies?
Keck et al. (2006)**	How effective is task-based interaction in fostering L2 grammar and vocabulary gains?
Russell and Spada (2006)**	Is oral and written corrective feedback effective?
Jeon and Kaya (2006)**	How effective is L2 pragmatics instruction, and does type of instruction matter?
Taylor et al. (2006)**	How effective is explicit reading strategy instruction in improving L2 reading comprehension?
Téllez and Waxman (2006)†	What are best teaching practices to teach English Language Learners in k-12 schools in the U.S.?
Thomas (2006)*	How do SLA researchers measure proficiency for research purposes in their studies, twelve years after Thomas (1994)?
Indefrey (2006)**	Are L1 and L2 processed differently, according to hemodynamic research evidence?
Mackey and Goo (2007)**	How effective is task-based interaction in fostering L2 grammar and vocabulary gains?
Truscott (2007)**	Is written corrective feedback effective for improving L2 writers' accuracy in new writing?
Abraham (2008)**	How effective are computer glosses in supporting L2 reading comprehension and L2 vocabulary learning?
Lee and Huang (2008)**	How effective is typographical input enhancement as an L2 grammar implicit teaching technique?
Spada and Tomita (2008)**	Are L2 instructional gains moderated by whether structures are simple or complex?

(*continued*)

Study	Key research question
Wa-Mbaleka (2008)**	What is the strength of the relationship between reading and vocabulary learning and how do different reading conditions and factors moderate that relationship?
In'nami and Koizumi (2009)**	Do multiple-choice response formats (when compared to open-ended response formats) lead to measurement differences in tests of L1 reading, L2 reading, or L2 listening?
Taylor (2009)**	How effective are L1 or L2 glosses in supporting L2 reading comprehension during computer-mediated L2 reading versus traditional (paper) L2 reading?
Jun, Ramirez and Cumming (2010)**	What type, focus, and amount of tutoring is beneficial for supporting literacy among adolescents, at what ages and from what language background?
Li (2010)**	How effective is L2 oral corrective feedback in laboratory, classroom, and small group settings?
Lyster and Saito (2010)**	How effective is L2 oral corrective feedback in classroom settings?
Oswald and Plonsky (2010)*	What range of magnitudes is typically found in applied linguistic meta-analyses and what benchmarks can be derived from these empirically attested magnitudes?
Spada and Tomita (2010)**	Are L2 instructional gains moderated by whether structures are simple or complex?
Adesope et al. (2011)	How effective are different instructional interventions designed to teach reading or writing to English ESL immigrant students in kindergarten through grade 6?
Biber et al. (2011)	How effective is written error correction on the development of writing proficiency and revision skills?
Plonsky (2011)**	How effective is L2 strategy training?
Plonsky and Gass (2011)*	What is the quality of the research methods employed in the domain of L2 oral interaction research over the last three decades?
Sauro (2011)*	What do we know about the potential of synchronous computer-mediated communication CMC for additional language learning given the research trends, methods, and findings in this research domain over the last two decades?
Sebastian et al. (2011)	How different or similar are the brain activation patterns of bilingual speakers when compared to first-language activation patterns of monolingual, and to what extent does this depend on whether the bilinguals exhibit high, moderate, or low levels of L2 proficiency?
Yoon (2011)*	How promising are learner corpus concordancing activities in cultivating linguistic awareness in L2 writing and learner autonomy in L2 writers?

(*continued*)

Study	Key research question
Yun (2011)**	How effective are text-only and text-plus-visual glosses to promote L2 vocabulary acquisition during online reading?
Hulstijn (2012)*	In what ways has language proficiency been measured across 140 studies published between 1998 and 2011 in one of the top journals in bilingualism?
Chiu (2013)**	How educationally effective is supporting L2 vocabulary learning through computer technologies?
Grgurović et al. (2013)**	How effective are language learning pedagogies in bringing about language outcomes when they are supported by computer technology, as compared to without technology?
Jackson and Suethanapornkul (2013)**	How much evidence is there for the predictions of the Cognition Hypothesis regarding increases in complexity, accuracy, and fluency when tasks are made cognitively more demanding?
Plonsky (2013)*	What are the trends, strengths, and weaknesses in the designs, statistical analyses, and reporting practices observed across 606 primary quantitative studies published from 1990 to 2010 in two top journals in the field of SLA?
Shintani et al. (2013)**	How effective is L2 instruction when it is comprehension based versus production based?
Jeon and Yamashita (2014)**	What are the typical association strengths observed between passage-level L2 reading comprehension and ten key reading variables that have been often investigated?
Koo, Becker and Kim (2014)**	Are some large-scale reading assessment items more consistently difficult for English Language Learners than their mainstream peers in grade 3 and grade 10?
Plonsky (2014)*	How have trends evolved from 1990 to 2010 in the designs, statistical analyses, and reporting practices reported across 606 primary quantitative studies, and what do these changes suggest about needed future improvements in quantitative L2 methods?

Note: *Systematic quantitative synthesis. **Meta-analysis. †Systematic qualitative synthesis. Studies are listed in chronological order. Full citations are given in the references.

Other bodies of literature, however, offer descriptive, qualitative or mixed-methods data that cannot be appraised exclusively by notions of effect or causality or they are too small for meta-analytic quantification. In such cases, synthesists employ a number of empirically oriented methods in order to systematically review the literature, but they do not necessarily engage in all the formal quantitative procedures of meta-analysis. It is for this reason

that we say all meta-analyses are also syntheses but not all syntheses are meta-analyses. It can be seen in Table 13.1 that in applied linguistics a good range of topics have been submitted to synthesis but mostly in the form of meta-analysis.

Underlying assumptions of the methodology

For a review to be considered a research synthesis, it must go beyond the collating and summarizing of individual study findings. Thus, none of the following summative techniques, in and of themselves, make a literature review into a synthesis: (a) presenting an authoritative narrative recount of past research activity in a given area, (b) summarizing many studies together in tabular form or (c) taking a count of how many studies yielded statistically significant findings in support of a given hypothesis. These are only traditional practices of good, traditional literature reviews. Research synthesists, on the other hand, must transcend such practices if they are to produce synthetic findings that are more than the sum of the parts and go beyond the individual results contained in any of the primary investigations synthesized (Norris & Ortega 2007).

Another trademark of many syntheses and meta-analyses (although not all) is a critical stance towards statistical significance. Many synthesists believe that while *probability* is an important dimension of quantitative research, *magnitude* is a distinct piece of information that is as important as probability and must be interpreted in conjunction with it (Kramer & Rosenthal 1999; Thompson 2006; for a recent discussion in L2 studies, see Plonsky 2014). When we pose questions about probability or statistical significance, we ask ourselves whether our observations are spurious (i.e., non-significant) or trustworthy (i.e., significant). This would only tell us whether the findings observed in a given study are likely to hold true if we carried out the same study again with a different sample drawn from the same population. Naturally, the issue of trustworthiness is very important in quantitative research paradigms, since it goes to the heart of generalizability. However, in the end, the questions of interest to experts – as well as to the public – are about magnitude. In applied linguistics, for example, an important question about the size of an effect would be: *How large or important* is the difference between providing instruction on an L2 grammar form or leaving it up to simple exposure to that form? And a central question about the strength of a relationship would be: If someone has below-average working memory capacity, *how severe or disadvantageous* can the consequences be for his or her potential to learn an L2 fast and well? By putting the emphasis on magnitude rather than probability or at least by always extracting and interpreting information about both probability and magnitude together, synthesists and meta-analysts seek to redress the many

misuses and abuses of statistical significance that have been documented in all the social sciences (Harlow et al. 1997; Ziliak & McCloskey 2008), including applied linguistics (Lazaraton 2000; Plonsky 2013, 2014).

Finally, and as can be surmised from the discussion thus far, research synthesis typically has a positivist and quantitative orientation. This orientation is expected, given that the approach arose out of the desire to make sense of quantitative findings (Glass 1976). Nevertheless, some synthesists have recognized the need to explore appropriate methods for the systematic review of accumulated *qualitative* literatures. For example, over twenty years ago, educational sociologists George Noblit and Dwight Hare developed an approach for synthesizing ethnographic research that they called meta-ethnography (Noblit & Hare 1988) and more recently, a research team in medical sociology led by Mary Dixon-Woods at Leicester University has explored new principles for conducting critical interpretive syntheses of research (see Dixon-Woods et al. 2007). In applied linguistics, as indicated in Table 13.1, only one systematic synthesis of qualitative research exists (Téllez & Waxman 2006). In the future, and mirroring the evolution of research synthesis in the wider landscape of the social sciences, interest in qualitative forms of systematic reviewing may grow in our field (Norris & Ortega 2007). Suri and Clarke (2009) offer a cogent discussion of the tensions and possibilities that surround the application of systematic research synthesis to qualitative research in the educational and social sciences.

Validity and trustworthiness in research synthesis

Several key issues impinge on the validity and trustworthiness of a research synthesis. Reliability and validity can be weakened or strengthened at the point in the research process when effect sizes need to be calculated, aggregated and reported, as you will see when you read about *Techniques and instruments for research syntheses* in a later section. In this section, I discuss important validity considerations that arise during the sampling of the primary studies to be included in the review.

Apples and oranges or the problem of relevance

From the beginning of the research process, the synthesist faces two basic questions of relevance that can affect the validity of the review findings: How closely related to a given research question must a study be in order to be included in the synthesis? and How similar to each other must studies be in order for their findings to be combined meaningfully in the same synthesis? These questions can only be answered by carefully considering

the purposes and research questions that guide a given synthesis. If we want to understand apples, it would be unwise to mix them with oranges in the same review. On the other hand, if we want to understand the two most popular fruits consumed daily in Western countries, then we may need to include precisely apples and oranges. If we want to understand fruit as a full category, moreover, we would want not only to mix apples and oranges but to include a wide palette of other types of fruit as well.

The apples and oranges question must be well justified in any synthesis and can only be satisfactorily answered if the synthesist has strong expert knowledge about the topic in question. In the end, therefore, substantive expertise in the research domain is as important as methodological expertise in the practice of research synthesis, and both are necessary to enhance the validity of the review. Coupled with the importance afforded to issues of coding, reliability and replicability, this is perhaps one of the reasons why research syntheses are often carried by teams rather than individuals (cf. Norris & Ortega 2006a).

Publication bias or the file-drawer problem

The problem of publication bias is little understood in the field of applied linguistics but well documented in all forms of quantitative research in the social sciences (Rothstein et al. 2005). First, studies that do not report at least some statistically significant result are likely to be rejected by journal referees. Secondly, authors are also inclined to give up on trying to publish a study which yielded no statistically significant outcomes, probably because they are well aware of the unspoken rejection bias of journals. In either case, the result is the same: studies that report statistically non-significant results are likely to be filed away in researchers' drawers and rarely make it into the public light. This means, in turn, that findings associated with a given research question are over-inflated if we only consider the universe of published studies, because statistically significant results are overrepresented (and statistically non-significant results underrepresented) in published literatures.

Therefore, good syntheses must always address the issue of publication bias. Among synthesists and meta-analysts, the preferred solution is to include unpublished or so-called fugitive literature in the synthesis. However, when including fugitive literature one must be doubly systematic and exhaustive in the sampling efforts. Particular care must be taken not to overrely on word-of-mouth knowledge of fugitive literature offered by immediate and far-away colleagues and their mentees and students. This knowledge, albeit valuable, is inevitably partial and most likely biased geographically and substantively. An arbitrary and incomplete sampling of unpublished studies can threaten the validity of the synthesis as much as an arbitrary and incomplete sampling of published studies would.

Whenever only published studies are included in a synthesis – unfortunately, the default case in applied linguistics thus far – the synthesist should assess the gravity of the publication bias at work. A wide range of sensitivity analyses has been developed specifically for this purpose (Rothstein et al. 2005). For example, Ross (1998) applied a mathematical estimation called the *fail-safe formula*; Norris and Ortega (2000) employed a visual technique called the *funnel plot*; and Li (2010) reported both the results of a funnel plot and a related technique called a *trim-and-fill analysis*. If both published and fugitive literatures are included in a synthesis, it is also informative to compare the main findings to the findings that obtain if published and fugitive studies are aggregated separately, as Taylor et al. (2006), Lee and Huang (2008), and Li (2010) did in our field.

Garbage in, garbage out or the problem of research quality

Ultimately, the quality of a synthesis is largely dependent on the quality of the primary evidence on which it is built. Some research synthesists follow the advice of educational psychologist Robert E. Slavin (1986), who advocated the *best evidence* approach to synthesizing. He proposed that only studies that meet the highest standards of methodological quality should be included in a synthesis. He was careful to warn that in the best evidence approach the task of determining what methodological rigour means must be explicitly rationalized anew by the synthesist for each research domain. It must also be justified for each study included in the synthesis, published or unpublished. A good illustration of these two points is offered in a meta-analysis by Slavin and Cheung (2005) that compared the effects of bilingual and English-only reading programs offered in elementary schools (and see also the online *Best Evidence Encyclopedia* of The Johns Hopkins University: http://www.bestevidence.org/, viewed 12 May 2014). Most synthesists, however, follow the broadly inclusive approach proposed by Gene Glass, the founder of meta- analysis (Glass 1976). As long as a study meets the set of inclusion criteria established at the onset of the synthesis, it will be included, so as to later perform sensitivity analyses that will help ascertain empirically whether, and in what ways, differences in research quality may have impacted the results of the review.

Norris and Ortega (2006b) argue, in agreement with most synthesists, that the inclusive approach serves well the field of applied linguistics. If the synthesist excludes certain studies a priori on grounds of poor research quality, this decision may always be contested, as methodological rigour is often in the eye of the beholder. Consider, for example, whether when carrying out a synthesis on effects of L2 instruction, intact classroom studies should be excluded because of their low internal validity due to lack of

experimental control or whether laboratory studies should be excluded for their low ecological validity, despite their high internal validity. Either decision would always dissatisfy one or another sector of the research domain. On the other hand, if the reviewer is as inclusive as it is reasonable at the initial stage of study sampling, the accumulated evidence can be made sense of more fully in the synthesis, and any possible biases introduced by the varying research quality of the individual designs can be more closely inspected and appraised.

Techniques and instruments for research synthesis

A research synthesis will typically entail several steps, which derive from the empirical perspective on reviewing it embraces:

- *Problem specification*: As a first step, the research problem or question to be synthesized must be specified carefully and precisely; this is analogous to the step, necessary in any primary study, of formulating well-thought-out research questions.
- *Literature search and study eligibility criteria*: These two steps entail explicating how the primary studies will be located and which ones will be included or excluded in the synthesis and why; this parallels the need in any empirical study, quantitative or qualitative, to carefully plan the selection of participants and justify sampling procedures.
- *Coding book development*: At this step, a systematic coding scheme must be devised by which all variables under study in the synthesis will be extracted from each primary report; this is analogous to the design of an instrument (e.g. a test) or a procedure (e.g. an interview or an observation protocol) that will elicit the relevant evidence from each participant in as consistent and well-motivated a fashion as possible.
- *Coding of studies*: This step documents how the study coding was done and how the reliability of the coding process was safeguarded.
- *Data analysis and display*: This step is parallel to the processes involved in analysing and displaying results in primary research. The evidence reported in study after study must be extracted first by application of the coding scheme and then it must be processed and made sense of; and subsequently, the findings must be organized in numerical, visual and narrative displays.

- *Interpretation and dissemination*: As with all research activities, synthesizing ultimately is a process that demands interpretive and dialogic efforts within a disciplinary community; the results must be interpreted in a historical and disciplinary context and the findings disseminated in reports that others can read, judge, replicate and use for their own purposes.

Many research syntheses, and all meta-analyses, involve the calculation of what is known as *effect sizes*. An effect size is an index that captures numerically, and in a standardized form, the strength of a relationship or the magnitude of a difference. Effect sizes must be demystified. We all know *r*, or the correlation coefficient, and have learned to interpret it. This is, in fact, an effect size. It expresses the strength of a relationship between two variables represented by two sets of scores. The closer *r* is to positive 1 or negative 1, the stronger the relationship between the two variables is known to be (and this observed relationship may or may not be trustworthy, depending on the output we obtain for the probability value or *p* associated with each *r*; see Phakiti this volume). Correlations are always interpretable in this same way (from 0 to plus or minus 1), regardless of the variety of tests, instruments and methods by which the scores submitted to correlations are obtained. There are also many instances when researchers report their findings in natural units, such as words produced, hours studied or cigarettes smoked. Proportions and percentages are also effect sizes that everyone understands and uses in daily matters. When this is the case, the results do not need to be translated into a mathematical unit of some other kind but can be most meaningfully expressed in natural units and in proportions and percentages. For example, according to the famous report released by the US Department of Health, Education, and Welfare in 1967, cigarette smoking has been proven to increase mortality rates at different bands of magnitude: when compared to no smoking at all, smoking up to half a pack a day (or less than ten cigarettes) increases the chance of mortality by 40 per cent, and if we raise the consumption of cigarettes by four times, to two packs or more a day, the risk rate increases by three times (120 per cent). These are effect sizes that everyone understands.

But when a researcher reports mean performances of groups and subsequently compares mean group differences, the scores reported are specific to that study's instruments (e.g. a 42-point difference on the TOEFL in one study, a 3.68 difference on a 15-point test in another study, and so on). In order to compare mean results across many studies, therefore, each of the individual study results must first be converted into a common index, or an effect size, which can then be aggregated together in a total mean effect size. This effect size is typically *d* (Cohen 1988), which simply expresses the difference between two group means in standard deviation units. For example, a *d* of 1.45 indicates that the experimental group scored 1.45 deviation units above the control group on the post-test or

that the advantage conferred by the treatment can be gauged to have had a magnitude of roughly one and a half deviation units in that study. Once we convert the results from each study into a *d*, we can average them and we will have the overall mean magnitude of a given treatment, based on however many studies we were able to combine. For example, a mean *d* of 1.45 indicates that, after inspecting study after study in our synthesis, we found that experimental groups had an average advantage over control groups of 1.45 deviation units in their favour.

It is important to emphasize that research synthesis, given its empirical take on the task of reviewing, adheres to the foundations of quantitative research in general. In order to enhance the reliability of the evidence and the validity of interpretations in primary quantitative research, it is important to consider the following issues against the context of the research questions and the design chosen: (a) sample size, (b) reliability of the instruments, (c) score distributions and (d) trustworthiness of the results. The same issues must also be considered by the synthesist, who will do well in thinking of each study as a 'participant' or 'informant' in the review. Thus, the synthesist must consider whether the number of accumulated studies is sufficient to carry out a full-blown meta-analysis or whether a synthesis using less powerful quantitative techniques is more appropriate in light of a small sample size. Furthermore, as with any elicitation instrument, the coding system employed to extract the same information consistently across all studies must be carefully developed and its reliable use enhanced and evaluated in the synthesis. In addition, if effect sizes are calculated, the synthesist should always make interpretations not only about the overall grand mean or average effect size but also about how spread out the results across individual studies are. Specifically, it is important to address the question: How representative of individual study results can a given average effect size be said to be? This can be achieved by reporting and inspecting the standard deviation associated with each mean effect size. Finally, whenever mean effect sizes are reported, confidence intervals that express the amount of error around the observed mean should also be calculated and reported. This information helps us determine how trustworthy the average findings really are, as it is equivalent to carrying out a statistical significance test.

Ethical considerations in research synthesis

Because synthesists only work with existing studies and previously reported results, they need not be concerned with norms for ethical conduct towards human participants. However, there are other ethical considerations worth mentioning. As with most quantitatively oriented research, syntheses and meta-analyses can be dangerously attractive and persuasive in their claim to take stock of accumulated evidence and in their aspiration to provide

so-called final answers to important but elusive questions in a given domain of study. A related danger is what well-known meta-analyst Robert Rosenthal has called high-tech statistication (Rosenthal & DiMatteo 2001), which occurs when technical virtuosity in quantitative synthetic techniques is achieved at the expense of substantive quality and depth.

Being cognizant of these traps and knowing the limits of research syntheses is important (Norris & Ortega 2006b, 2007). No single synthesis can give a definitive answer to a research problem, because research is a human enterprise that is contingent on the time and space in which it is produced. Like all research, syntheses may be able to answer questions of now and here, but these questions will evolve. All knowledge will always be re-evaluated and re-calibrated, as human history and consciousness change. In addition, no dose of statistical or technical expertise can make up for lack of substantive expertise. Finally, research syntheses are always purely descriptive and correlational and cannot completely dispel debates surrounding causality, because the synthesist has no choice over the studies that make it into the review or over how individual researchers operationalized their variables and designed their investigations. In sum, syntheses can only produce evidence that is firmly rooted in the contexts in which the primary research has been carried out.

In the end, an ethical approach to practising research synthesis may come from 'a research ethic to think and act synthetically', and from 'a commitment to intelligible and respectful communication: between past and present studies, across theory and research and practice divides, between research contexts or camps, and among members of each research community' (Norris & Ortega 2006b, p. 36).

A sample study

I conclude this chapter with a quick tour of a meta-analysis that my colleague John Norris and I carried out (Norris & Ortega 2000) and which illustrates the methodology and its main concepts and procedures.

Problem specification

We set out to investigate the effectiveness of different types of L2 instruction. After examining the concepts most often discussed in this research domain, we decided against trying to gauge the effects of particular pedagogical techniques (e.g. recasts, grammar translation, input processing) because there were insufficient numbers of studies accumulated for any such technique. On the other hand, we adopted the theoretical notions of focus on form versus forms (Long 1991) and explicit versus implicit grammar

teaching (DeKeyser 1995), because we reasoned they were workable characterizations of type of instruction that could be found in all specific techniques. We also decided to include as moderating variables the type of outcome measure (e.g. grammaticality judgement or free production task), the length of instruction (e.g. half an hour or 50 hours) and the durability of effects (i.e., whether any gains were maintained on delayed post-tests). The rationale was that these variables had been discussed in previous literature as important concerns when evaluating the effectiveness of L2 instruction and that information for at least the first two moderating variables could be extracted from each study (we expected that delayed post-tests would be present in only a subset of the studies).

Literature search and study eligibility criteria

Although we initially identified over 250 studies that were potentially relevant via electronic and manual searches, subsequently only seventy-seven study reports met the inclusion criteria for the synthesis that we had previously developed based on our knowledge of the research domain. Furthermore, we were able to include only a subset of these (specifically, forty-nine unique sample studies) in the meta-analysis part of the synthesis, because the remainder studies did not contain sufficient information to calculate effect sizes. The studies had been published between 1980 and 1998 and therefore represented the research domain as was practised in the 1980s and 1990s. The designs varied widely but all were (quasi-) experimental and involved instruction of a specific L2 form as the independent variable and measurement of performance on the same specific form as the dependent variable(s). After long deliberations, we decided to include only published studies and to inspect publication bias by means of a funnel plot. While we found no direct evidence of publication bias, we cautioned that this was most likely due to the fact that most individual researchers used complex multiple-treatment designs and therefore were able to report both statistically significant and non-significant results for different treatments within the same study. We also decided not to use methodological quality as a criterion for excluding studies and instead to adopt a broadly inclusive approach. That is, we decided to investigate research quality as an empirical matter, inspecting and synthesizing in full detail the research practices found across all seventy-seven studies.

Coding book development

An important and time-consuming step was to develop the coding book, which had to include a large number of categories in order to address each research question. For example, we decided that each study would be coded

for methodological and background features such as learner characteristics, study design, sample size, length of treatment, timing of tests and statistical information reported. We also coded each study (and for multiple-treatment designs, each treatment group within each study) for substantive features, including: type of instruction (with five values: focus on form, focus on forms, focus on meaning, explicit and implicit), type of outcome measure (with four values: meta-linguistic judgement, selected response, constrained constructed response and free constructed response) and length of instruction (with four values: brief, short, medium and long; these categories were defined bottom-up, according to the range of observed instructional lengths in the seventy-seven studies).

Coding of studies

Both of us were involved as coders in the process of study coding. For most methodological, low-inference categories (e.g. year and type of publication) each coded a different half of the studies. Because we knew type of instruction and type of outcome measure would involve high-inference coding decisions, we decided to calculate and report reliability for these two categories only. Therefore, we set apart 20 per cent of the sample of studies and independently coded them for these substantive features. The reliability of the codings was satisfactory with values above 0.90 for both simple per cent agreement and *Cohen's kappa* (see Phakiti this volume).

Data analysis and display

We tallied all values for the coded methodological study features, aggregated them and presented them in tabular form (these results can be found in Tables 1–6 and Figures 1 and 2 in the original 2000 report). In order to answer the main research questions, we extracted effect sizes from the forty-nine unique sample studies for which sufficient information for this calculation was available. It is important to note that we calculated two distinct types of effect sizes, which we then aggregated and reported separately (as it should always be done if both types of effect sizes are used in the same meta-analysis): (a) standardized mean difference *d*s, which compared post-test performances of treatment versus control groups and (b) standardized mean gain *d*s, also called within-group pre-post contrasts, which expressed pre-to-post-test change for each group, including control or baseline groups. For each of the two effect size types, aggregations resulted in average effect sizes. We presented the information in the form of means, standard deviations and confidence intervals (which can be found in Tables 7–12 and Figures 3–8 in the original 2000 report).

Interpretation and dissemination

We found that L2 instruction overall was superior to simple L2 exposure or meaning-driven communication by nearly a full standard deviation unit on average (d = 0.96 based on forty-nine studies) and that explicit treatments were clearly superior to implicit treatments, in terms of both magnitude (d = 1.13 versus d = 0.54, based on sixty-nine and twenty-nine contrasts, respectively) and probability (the confidence intervals around these two means did not overlap, which amounts to saying that the two means were statistically significantly different from each other). On the other hand, the difference between focus on form and focus on forms treatments was small (d = 1.00 versus 0.93) and statistically not significant, indicating that both qualities of instruction were effective and neither one was superior to the other in the evidence examined (forty-three and fifty-five contrasts, respectively). We could offer no conclusive answers with regard to the influence of type of outcome measure or the varying lengths of instruction, due to insufficient sample size across categories for both variables. However, to our surprise, we found that the twenty-two studies which featured delayed post-tests yielded an average effect size of 1.02 and that the lower boundary of the associated confidence interval was 0.78, and from this we concluded that the effects of instruction were indeed durable for the participants involved in these twenty-two studies. Finally, we documented empirically a number of endemic methodological weaknesses typical of this domain, based on the full sample size of seventy-seven studies. We discussed these problems in detail and proposed some solutions.

It is important to stress that the main findings of this synthesis, as much as any other one, are contingent upon and grounded in the available accumulated evidence (Norris & Ortega 2006b). As authors of the synthesis, we have always felt readers of this study should not think of these results as directly generalizable to the abstract concept of 'L2 instruction'. This is simply because the real world of L2 instruction goes well beyond the world captured in those seventy-seven (or forty-nine) studies and offers a much richer kaleidoscope of contexts for the instruction of additional languages and much more varied approaches to the teaching of grammar than was possibly investigated in the particular studies synthesized. Moreover, the many methodological weaknesses uncovered and carefully documented also call for caution when extrapolating the findings beyond the concrete body of evidence synthesized. More modestly, we find value in the results of this meta-analysis because we hope they offer 'a useful empirical context within which future single-study findings from L2 type-of-instruction research can be more meaningfully interpreted' (Norris & Ortega 2000, pp. 499–500).

Resources for further reading

Cooper, H, Hedges, LV & Valentine, JC (eds), 2009, *The Handbook of Research Synthesis and Meta-Analysis*, 2nd edn, Russell Sage Foundation, New York, NY.

This edited collection constitutes the most comprehensive and encyclopedic treatment of meta-analysis to date. The utility (and difficulty) of chapters varies, but all have been written by international experts in specialized sub-areas of meta-analysis.

Hunt, M 1997, *How Science Takes Stock: The Story of Meta-Analysis*, Russell Sage Foundation, New York, NY.

This book offers a lively and accessible chronicle of the history of meta-analysis. Readers can learn a great deal of technical concepts in an intuitive fashion thanks to the variety of concrete examples offered from the fields of psychology, education and medicine.

Light, R & Pillemer, D 1984, *Summing Up: The Science of Reviewing Research*, Harvard University Press, Cambridge, MA.

This book is an unsurpassed classic treatise about research synthesis. It goes well beyond meta-analysis and offers many visual and descriptive techniques for synthesizing quantitative findings. For this reason, it is an invaluable tool to learn about the many options available in the methodology of synthesis.

Lipsey, MW & Wilson, DB 2001, *Practical Meta-Analysis*, Sage, Thousand Oaks, CA.

There are several textbook-like manuals about meta-analysis, but this one is particularly accessible and complete. It answers most technical questions that beginning meta-analysts will have about formulas for the calculation of different types of effect sizes, strategies for keeping track of study codings, and so on.

Cumming, G 2011. *Understanding the New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis*, Routledge, New York, NY.

This statistics textbook requires advanced background knowledge in statistics and will not be easy to read for the average applied linguist. However, it is well worth the extra investment, as it can teach researchers in second language studies a great deal about how to think outside the box when it comes to inferential statistics and the relationships among numbers, research and meta-analysis.

Norris, JM & Ortega, L (eds), 2006, *Synthesizing Research on Language Learning and Teaching*, John Benjamins, Amsterdam.

This is the first and thus far only collection that exemplifies the methodology of research synthesis in applied linguistics. It includes applications of the methodology to universal grammar, interaction, negative feedback, pragmatics, reading strategies, the measurement of proficiency and best practices for English language learners in US schools. The variety of topics sampled in the empirical studies will help readers understand how synthesis can be applied to very diverse areas of research within applied linguistics. In the first chapter, the co-editors offer a critical, extensive overview of the principles and uses of synthesis.

References

Abraham, L 2008, 'Computer-mediated glosses in second language reading comprehension and vocabulary learning: A meta-analysis', *Computer Assisted Language Learning*, vol. 21, no. 3, pp. 199–226.

Adesope, OO, Lavin, T, Thompson, T & Ungerleider, C 2011, 'Pedagogical strategies for teaching literacy to ESL immigrant students: A meta-analysis', *British Journal of Educational Psychology*, vol. 81, no. 4, pp. 629–653.

Biber, D, Nekrasova, T & Horn, B 2011, 'The effectiveness of feedback for L1-English and L2-writing development: A meta-analysis' *TOEFL iBT™ Research Report*, 14, viewed 14 February 2014, http://www.ets.org/Media/Research/pdf/RR-11-05.pdf.

Chiu, Y-H 2013, 'Computer-assisted second language vocabulary instruction: A meta-analysis', *British Journal of Educational Technology*, vol. 44, no. 2, pp. E52–E56.

Cohen, J 1988, *Statistical Power Analysis for the Behavioral Sciences*, 2nd edn, Lawrence Erlbaum, Hillsdale, NJ.

DeKeyser, R 1995, 'Learning second language grammar rules: An experiment with a miniature linguistic system', *Studies in Second Language Acquisition*, vol. 17, no. 3, pp. 379–410.

Department of Health, Education, and Welfare 1967, *The Health Consequences of Smoking: A Public Health Service Review*, Public Health Service Publication No. 1696, Washington, DC.

Dinsmore, TH 2006, 'Principles, parameters, and SLA: A retrospective meta-analytic investigation into adult L2 learners' access to universal grammar', in J M Norris & L Ortega (eds), *Synthesizing Research on Language Learning and Teaching*, John Benjamins, Amsterdam, pp. 53–90.

Dixon-Woods, M, Booth, A & Sutton, AJ 2007, 'Synthesising qualitative research: A review of published reports', *Qualitative Research*, vol. 7, no. 3, pp. 375–422.

Glass, GV 1976, 'Primary, secondary, and meta-analysis of research', *Educational Researcher*, vol. 5, no. 10, pp. 3–8.

Goldschneider, J & DeKeyser, RM 2001, 'Explaining the "natural order of L2 morpheme acquisition" in English: A meta-analysis of multiple determinants', *Language Learning*, vol. 51, no. 1, pp. 1–50.

Grgurović, M, Chapelle, CA & Shelley, MC 2013. 'A meta-analysis of effectiveness studies on computer technology-supported language learning', *ReCALL*, vol. 25, no. 2, pp. 165–198.

Harlow, L, Mulaik, S & Steiger, J (eds), 1997, *What If There Were No Significant Tests?*, Lawrence Erlbaum, Mahwah, NJ.

Hulstijn, J. H. 2012. 'The construct of language proficiency in the study of bilingualism from a cognitive perspective', *Bilingualism: Language and Cognition*, vol. 15, no. 2, pp. 422–433.

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*, vol. 26, no. 2, pp. 219–244.

Indefrey, P 2006, 'A meta-analysis of hemodynamic studies on first and second language processing: Which suggested differences can we trust and what do they mean?', *Language Learning*, vol. 56, no. s1, pp. 279–304.

Jackson, D & Suethanapornkul, S 2013. 'The cognition hypothesis: A synthesis and meta-analysis of research on second language task complexity', *Language Learning*, vol. 63, no. 2, pp. 330–367.
Jeon, EH & Kaya, T 2006, 'Effects of L2 instruction on interlanguage pragmatic development: A meta-analysis', in JM Norris & L Ortega (eds), *Synthesizing Research on Language Learning and Teaching*, John Benjamins, Amsterdam, pp. 165–211.
Jeon, EH & Yamashita, J 2014, 'L2 reading comprehension and its correlates: A meta-analysis', *Language Learning*, vol. 64, no. 1, pp. 160–212.
Jun, SW, Ramirez, G & Cumming, A 2010. 'Tutoring adolescents in literacy: A meta-analysis', *McGill Journal of Education*, vol. 45, no. 2, pp. 219–238.
Keck, CM, Iberri-Shea, G, Tracy-Ventura, N & Wa-Mbaleka, S 2006, 'Investigating the empirical link between task-based interaction and acquisition: A meta-analysis', in JM Norris & L Ortega (eds), *Synthesizing Research on Language Learning and Teaching*, John Benjamins, Amsterdam, pp. 91–131.
Koo, J, Becker, BJ & Kim, Y-S 2014. 'Examining differential item functioning trends for English language learners in a reading test: A meta-analytical approach', *Language Testing*, vol. 31, no. 1, pp. 89–109.
Kramer, SH & Rosenthal, R 1999, 'Effect sizes and significance levels in small-sample research', in RH Hoyle (ed.), *Statistical Strategies for Small Sample Research*, Sage, Thousand Oaks, CA, pp. 59–79.
Lazaraton, A 2000, 'Current trends in research methodology and statistics in applied linguistics', *TESOL Quarterly*, vol. 34, no. 1, pp. 175–181.
Lee, S-K & Huang, HT 2008, 'Visual input enhancement and grammar learning: A meta-analytic review', *Studies in Second Language Acquisition*, vol. 30, no. 3, pp. 307–331.
Li, S 2010, 'The effectiveness of corrective feedback in SLA: A meta-analysis', *Language Learning*, vol. 60, no. 2, pp. 309–365.
Long, MH 1991, 'Focus on form: A design feature in language teaching methodology', in KD Bot, R Ginsberg & C Kramsch (eds), *Foreign Language Research in Cross-Cultural Perspective*, John Benjamins, Amsterdam, pp. 39–52.
Lyster, R & Saito, K 2010. 'Oral feedback in classroom SLA: A meta-analysis', *Studies in Second Language Acquisition*, vol. 32, no. 2, pp. 265–302.
Mackey, AV, Goo, JM 2007, 'Interaction research in SLA: A meta-analysis and research synthesis', in A Mackey (ed.), *Conversational Interaction in Second Language Acquisition*, Oxford University Press, New York, NY, pp. 379–452.
Masgoret, A-M & Gardner, RC 2003, 'Attitudes, motivation, and second language learning: A meta-analysis of studies conducted by Gardner and associates', *Language Learning*, vol. 53, no. s1, pp. 123–163.
Noblit, GW & Hare, RD 1988, *Meta-Ethnography: Synthesizing Qualitative Studies*, Sage, Newbury Park, CA.
Norris, JM & Ortega, L 2000, 'Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis', *Language Learning*, vol. 50, no. 3, pp. 417–528.
Norris, JM & Ortega, L (eds), 2006a, *Synthesizing Research on Language Learning and Teaching*, John Benjamins, Amsterdam.
——— 2006b, 'The value and practice of research synthesis for language learning and teaching', in JM Norris & L Ortega (eds), *Synthesizing Research on Language Learning and Teaching*, John Benjamins, Amsterdam, pp. 3–50.

——— 2007, 'The future of research synthesis in applied linguistics: Beyond art or science', *TESOL Quarterly*, vol. 41, no. 4, pp. 805–815.
Ortega, L 2003, 'Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing', *Applied Linguistics*, vol. 24, no. 4, pp. 492–518.
Oswald, FL & Plonsky, L 2010. 'Meta-analysis in second language research: Choices and challenges', *Annual Review of Applied Linguistics*, vol. 30, pp. 85–110.
Plonsky, L 2011. 'The effectiveness of second language strategy instruction: A meta-analysis', *Language Learning*, vol. 61, no. 4, pp. 993–1038.
——— 2013, 'Study quality in SLA: An assessment of designs, analyses, and reporting practices in quantitative L2 research', *Studies in Second Language Acquisition*, vol. 35, no. 4, pp. 655–687.
——— 2014, 'Study quality in quantitative L2 research (1990–2010): A methodological synthesis and call for reform', *The Modern Language Journal*, vol. 98, no. 1, pp. 450–470.
Plonsky, L & Gass, SM 2011, 'Quantitative research methods, study quality, and outcomes: The case of interaction research', *Language Learning*, vol. 61, no. 2, pp. 325–366.
Roessingh, H 2004, 'Effective high school ESL programs: A synthesis and meta-analysis', *Canadian Modern Language Review*, vol. 60, no. 5, pp. 611–636.
Rosenthal, R & DiMatteo, MR 2001, 'Meta-analysis: Recent developments in quantitative methods for literature reviews', *Annual Review of Psychology*, vol. 52, pp. 59–82.
Ross, S 1998, 'Self-assessment in second language testing: A meta-analysis and analysis of experiential factors', *Language Testing*, vol. 15, no.1, pp. 1–20.
Rothstein, HR, Sutton, AJ & Borenstein, M (eds), 2005, *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments*, Wiley, Chichester.
Russell, J & Spada, N 2006, 'The effectiveness of corrective feedback for the acquisition of L2 grammar: A meta-analysis of the research', in JM Norris & L Ortega (eds), *Synthesizing Research on Language Learning and Teaching*, John Benjamins, Amsterdam, pp. 133–164.
Sauro, S 2011, 'SCMC for SLA: A research synthesis', *CALICO Journal*, vol. 28, no. 2, pp. 369–391.
Sebastian, R, Laird, AR & Kiran, S 2011, 'Meta-analysis of the neural representation of first language and second language', *Applied Psycholinguistics*, vol. 32, no. 4, pp. 799–819.
Shintani, N, Li, S & Ellis, R 2013, 'Comprehension-based versus production-based instruction: A meta-analysis of comparative studies', *Language Learning*, vol. 63, no. 2, pp. 296–329.
Slavin, RE 1986, 'Best evidence synthesis: An alternative to meta-analytic and traditional reviews', *Educational Researcher*, vol. 15, no. 9, pp. 5–11.
Slavin, RE & Cheung, A 2005, 'A synthesis of research on language of reading instruction for English language learners', *Review of Educational Research*, vol. 75, no. 2, pp. 247–284.
Spada, N & Tomita, Y 2008, 'The complexities of selecting complex (and simple) forms in instructed SLA research', in A Housen & F Kuiken (eds), *Proceedings of the Complexity, Accuracy and Fluency (CAF) Conference*, University of Brussels, Brussels, pp. 227–254.

——— 2010, 'Interactions between type of instruction and type of language feature: A meta-analysis', *Language Learning*, vol. 60, no. 2, pp. 263–308.

Suri, H & Clarke, D 2009, 'Advancements in research synthesis methods: From a methodologically inclusive perspective', *Review of Educational Research*, vol. 79, no. 1, pp. 395–430.

Taylor, A, Stevens, JR & Asher, JW 2006, 'The effects of explicit reading strategy training on L2 reading comprehension: A meta-analysis', in JM Norris & L Ortega (eds), *Synthesizing Research on Language Learning and Teaching*, John Benjamins, Amsterdam, pp. 213–244.

Taylor, AM 2009, 'CALL-based versus paper-based glosses: Is there a difference in reading comprehension?', *CALICO Journal*, vol. 27, no. 1, pp. 147–160.

Téllez, K & Waxman, HC 2006, 'A meta-synthesis of qualitative research on effective teaching practices for English Language Learners', in JM Norris & L Ortega (eds), *Synthesizing Research on Language Learning and Teaching*, John Benjamins, Amsterdam, pp. 245–277.

Thomas, M 1994, 'Assessment of L2 proficiency in second language acquisition research', *Language Learning*, vol. 44, no. 2, pp. 307–336.

——— 2006, 'Research synthesis and historiography: The case of assessment of second language proficiency', in JM Norris & L Ortega (eds), *Synthesizing Research on Language Learning and Teaching*, John Benjamins, Amsterdam, pp. 279–298.

Thompson, B 2006, *Foundations of Behavioral Statistics: An Insight-Based Approach*, Guilford, New York, NY.

Truscott, J 2007, 'The effect of error correction on learners' ability to write accurately', *Journal of Second Language Writing*, vol. 16, no. 4, pp. 255–272.

Wa-Mbaleka, S 2008, *A Meta-Analysis Investigating the Effects of Reading on Second Language Vocabulary Learning*, VDM Verlag, Saarbrücken.

Yoon, C 2011, 'Concordancing in L2 writing class: An overview of research and issues', *Journal of English for Academic Purposes*, vol. 10, no. 3, pp. 130–139.

Yun, J 2011, 'The effects of hypertext glosses on L2 vocabulary acquisition: A meta-analysis', *Computer Assisted Language Learning*, vol. 24, no. 1, pp. 39–58.

Ziliak, ST & McCloskey, DN 2008, *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives*, The University of Michigan Press, Ann Arbor, MI.