

Another Look at Norris and Ortega (2000)

Hye Won Shin¹

Teachers College, Columbia University

ABSTRACT

Since Norris and Ortega's (2000) seminal work on the effectiveness of second language instruction, there has been a proliferation of meta-analyses in the field of applied linguistics. Subsequent meta-analysts, however, have uncritically followed the methodological choices made by Norris and Ortega. This paper suggests a critical reevaluation of the methodological procedures underlying the Norris and Ortega (2000) meta-analysis. I reexamined their procedures, and reassessed the 49 unique samples they used in their meta-analysis. In doing so, I identified three key methodological limitations with the study, pertaining, respectively, to (a) the data collection procedure, (b) the coding system, and (c) the statistical analysis. I argue that the lack of data quality inherent in the primary studies, the oversimplified coding scheme, and the inappropriate use of effect size statistics combine to compromise the validity of the conclusions Norris and Ortega have drawn from their meta-analysis. I subsequently provide alternative procedures which may yield a more empirically sound research synthesis, recommending, for future meta-analysts, the 'best evidence synthesis' approach where conclusions are drawn from combining quantitative and qualitative analyses.

INTRODUCTION

Broadly defined, "meta-analysis" (Glass, 1976) or "research synthesis" (Cooper & Hedges, 1994) refers to a method for combining the results of multiple primary studies in a particular research domain. The idea is to integrate the empirical findings to arrive at an overall conclusion across these studies, quantitatively summarizing a large body of literature (Green & Hall, 1984).² In this sense, meta-analysis can be viewed as a *quantitative literature review* (Dörnyei, 2007).

Norris and Ortega's (2000) study, "Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis," analyzes previous empirical studies that attempt to explore the interface of second language pedagogy and second language acquisition (SLA). While the role of instruction in promoting L2 development has by and large been established following Long's (1983) findings, and later Doughty and Williams' (1998) work, Norris and Ortega's research synthesis has made notable headway in second language research. Specifically, they have provided researchers with a macroscopic view of the effectiveness of L2 instruction and have helped practitioners in identifying which instructional practices facilitate second language

¹ Hye Won Shin is a doctoral student in Applied Linguistics at Teachers College, Columbia University. She is currently engaged in preparing for research on the efficacy of written corrective feedback. Correspondence should be sent to Hye Won Shin, 5 Owen St., Unit B, Westbury, NY 11590. Email: hs2509@columbia.edu.

² Other methods for aggregating a large number of studies include vote-counting, narrative review, and accumulation of p-values across studies. See Hunter et al., (1982) for a further review of these approaches.

learning. More importantly, their study has set the very first example of how meta-analysis could be adapted to the field, and consequently, later researchers have produced research syntheses that are similar to Norris and Ortega.

Despite the significant contributions mentioned, the synthesis by Norris and Ortega (2000) exhibits some weaknesses. This paper aims to critically evaluate the methodological procedures employed by Norris and Ortega and propose alternative procedures for future research syntheses. In the ensuing sections, Norris and Ortega's synthesis will first be reviewed. Then, the next section will call attention to the key limitations in Norris and Ortega's methodological procedures. Finally, implications and future research directions will be discussed.

OVERVIEW OF NORRIS & ORTEGA'S (2000) META-ANALYSIS

The main objective of Norris and Ortega's (2000) synthesis was twofold: (1) to provide a quantitative summary of empirical findings about L2 instruction, and (2) to evaluate the research methods and reporting practices that had led to these findings.

Accordingly, six research questions were posed in the study (pp. 428-429):

- a. How effective is L2 instruction overall and relative to simple exposure or meaning-driven communication?
- b. What is the relative effectiveness of different types and categories of L2 instruction?
- c. Does type of outcome measure influence observed instructional effectiveness?
- d. Does length of instruction influence observed instructional effectiveness?
- e. Does instructional effect last beyond immediate postexperimental observations?
- f. To what extent has primary research provided answers to these questions?

Norris and Ortega (2000) began their data collection for the meta-analysis through an extensive literature search, locating all possible studies within second language instruction, the targeted domain. The first pass through the literature derived 250-plus studies primarily through the Educational Resources Information Center (ERIC) electronic database. In addition, they utilized other search techniques including reviewing back issues of 14 academic journals, investigating citations from review sections, and cross-checking reference sections to retrieve more potentially relevant studies.³ However, they excluded "fugitive" studies such as "unpublished papers, dissertations and theses, [and] conference presentations" (p. 431).

Next, Norris and Ortega (2000) used the following set of inclusion criteria: (1) studies were to be published between 1980 and 1998, (2) studies were to have a quasi-experimental or experimental design, (3) studies' independent measures were to include an instructional treatment, and targeted forms or functions in morphology, syntax, or pragmatics, and (4) studies' dependent measures were to include quantitative measures of language behavior. The criteria yielded a total of 77 studies. Each study was further reviewed for substantive and methodological features, respectively. Of the 77 studies, 49 unique samples that were presented in 45 studies

³ The 14 academic journals include: *Applied Linguistics*, *Applied Psycholinguistics*, *Applied Language Learning*, *Canadian Modern Language Review*, *Foreign Language Annals*, *JALT Journal*, *Language Learning*, *Language Teaching Research*, *The Modern Language Journal*, *RELC Journal*, *Second language Research*, *Studies in Second Language Acquisition*, *System*, and *TESOL Quarterly*.

were deemed to present sufficient data⁴ for inclusion in the meta-analysis. The resultant meta-analysis summarized empirical findings from those 49 experimental and quasi-experimental data samples.

For the data samples collected, Cohen's d was employed to measure the effect size as shown by the formula $d = \frac{\bar{X}_1 - \bar{X}_2}{S_p}$, $S_p = \sqrt{\frac{(n_1 - 1)s_1 + (n_2 - 1)s_2}{(n_1 - 1) + (n_2 - 1)}}$. If descriptive statistics were not available, Rosenthal's (1994) formulae (as cited in Norris and Ortega, 2000, pp. 443-444), described below, were used to estimate the effect size: $d = \frac{t(n_1 + n_2)}{(\sqrt{df_{error}})(\sqrt{n_1 n_2})}$, or $d = \frac{(\sqrt{F})(n_1 + n_2)}{(\sqrt{df_{error}})(\sqrt{n_1 n_2})}$. Effect sizes were pooled across the 49 data samples to quantitatively summarize the findings.

Overall, Norris and Ortega's (2000) results suggest that (1) explicit instruction is more effective than implicit instruction, (2) Focus on Form (FonF) and Focus on Forms (FonFS) are equally effective, and (3) instructional effectiveness has an extended duration effect that lasts beyond post-experimental observations in L2 learning.

THREE METHODOLOGICAL ISSUES

An in-depth review of the 45 primary studies and the methodological procedures conducted by Norris and Ortega (2000) reveals three methodological issues that deserve further investigation. These issues are related to: (a) the data collection procedure, (b) the coding system, and (c) the statistical analysis.

Data Collection Issues

Lack of Randomization in Primary Studies

A fundamental principle in any experimental design is the idea of *randomization* – the likelihood of subjects randomly assigned to experimental conditions such that each subject has the same probability of being assigned to given conditions (Kerlinger, 1986). Randomization allows us to infer that an experimental treatment has a causal effect. However, should the treatment effect be confounded with other factors because of a lack of randomization, the internal validity of a study may be threatened. In other words, since an alternative explanation for the observed relationships cannot be ruled out, one cannot say that the result is solely due to the treatment effect. Accordingly, the result from poorly randomized experiment is neither replicable nor generalizable. In Kerlinger's words, "It is not possible to overrate the importance of both the idea and the practical measures that come from it [randomization] to improve experimentation and inference" (pp. 111-115).

The principle of randomization should not be taken lightly in the collection of data samples for a meta-analysis. A data sample in meta-analysis can be defined as the quantitative finding(s) from a primary study. The average effect size, then, is calculated by using the effect sizes obtained from the data samples. In this sense, randomization is crucial to data sample

⁴ For example, one study may have two or more sub-studies. Norris and Ortega (2000) regard those sub-studies as independent data samples.

collection, given the underlying assumption that the individual effect size reflects true treatment effect after controlling for all possible confounding effects. Otherwise, the average effect size will be of little value.

However, the principle of randomization is not evident in the primary studies chosen by Norris and Ortega (2000). To achieve randomization, experimental designs employ random assignment of subjects to treatment and control groups, and this ensures that each participant has the same opportunity to be assigned to either treatment or control group. Surprisingly, 47% (i.e., 23 out of the 49) of data samples used by Norris and Ortega did not report that they used random assignment. Acknowledging the possibility of confounding effects in the absence of randomization, Alanen (1995) states: "... several factors that may have affected the results of this experiment were not controlled for, including the level of proficiency achieved in the languages studied, naturalistic language learning experience, and language aptitude" (p. 266).

Note that in a true random design, subjects should be assigned non-systematically or randomly first, and then a pretest should be performed to assure that random assignment is indeed well done. When random assignment was not employed, however, researchers can still use a pretest to make sure that experimental and control groups are not different from one another in at least one dimension; often, this is the proficiency level of the L2 learners. Of the 23 data samples without random assignment, 17 samples performed a pretest. For example, Scott (1989) found statistically significant differences between the groups ($p < 0.05$), in which a morning class was assigned to one group, and an afternoon class was assigned to the other group. In order to account for initial group differences, once detected through the pretest, researchers can rely on statistical techniques such as ANCOVA or repeated measures ANOVA. Some of the studies, however, did not report any statistical test results for the pretest (e.g., Ellis, Rosszell, & Takashima, 1994; Scott, 1990), raising questions about the validity of their findings.

Another issue regarding randomization is the *sample size*. The sample size itself is critical for the randomization of primary studies, and in turn, for the validity of the research synthesis. When the group sample sizes are small, it becomes increasingly difficult to assess whether or not the sample was drawn from a normally distributed population. In other words, if the sample size is too small, the random assignment procedure, even if used, cannot guarantee randomization. One common observation in the primary studies used by Norris and Ortega (2000) is that the sample size is rather small: 28.2% consist of sample sizes of 10 or less and 88.5% of samples sizes of 30 or less. A case in point is the study by Jourdenais, Ota, Stauffer, Boyson, and Doughty (1995), which investigated the effects of textual enhancement on language development. The study consists of 10 subjects that were either assigned to the treatment or the comparison group. Jourdenais et al. state: "Because the participants had been carefully matched ... and randomly assigned ..., it would be safely assumed that the two groups were equal in ability and that any differences that obtained would be due to the enhancement procedure" (p. 200). For studies with just 10 participants, however, random assignment is likely to be of little value. In fact, the two worst performing participants were both placed in the treatment group. This may have biased the outcome given that 40% (2 out of 5) of the subjects in one group are considered to be 'bad apples.' Notwithstanding a systematic matching of participants based on pretest, a replication of this study is unlikely to produce the same results. If we sample 10 students and randomly assign them to two groups, this time we may have the two worst performers in the control group, and this might yield quite different findings from the ones of Jourdenais et al. Despite the researchers' claim, the veracity of random assignment in this study

is questionable, and thus, findings from this study may not be reliable. This understanding is in accord with the conceptualization of the randomization.

Another example is the study by Nagata (1997a), which examined the effectiveness of computer-assisted metalinguistic instruction for teaching grammatical features. The researcher randomly assigned 14 subjects into two groups, and reported that there was no difference between the two groups at 95% significance level. The reported t-statistic is rather large (-2.00), implying that there might have been a significant difference between the groups if a more conservative critical value (i.e., 90% significance level or $p < 0.10$) was used. Note that the null hypothesis being tested for a pretest is that there *is* a difference between the groups even after random assignment. This large t-statistic is likely due to the presence of small samples, i.e., 7 subjects per group. Interestingly, another study by Nagata (1997b) adopted the same procedure, assigning 30 subjects into two groups. The t-statistic from the pretest was 0.55 ($p=0.59$), suggesting that the same random assignment procedure yielded more homogeneous groups, which lessens the concern about randomization. Nagata's two studies illustrate the importance of the sample size when selecting a data sample for meta-analysis, even when random assignment is performed.

The aforementioned observations raise the question of whether the principle of randomization has received due attention in Norris and Ortega's (2000) synthesis. In fact, Norris and Ortega explicitly state: "no such decisions [with respect to quality criteria for inclusion decisions] were made based on the validity of the primary research reported" (p. 434). Therein lies the fundamental weakness of the Norris and Ortega's meta-analysis. When meta-analysts summarize effect sizes from data samples that lack randomization, the resulting average effect size is almost certain to be confounded. They must also establish firm criteria as to what constitutes a reliable data sample. As Dörnyei (2007) points out, "the quality of the analysis ultimately depends on the underlying studies" (p. 241).

Instrument Validity and Coding Consistency in Primary Studies

Another consideration is the validity of the instruments used in the primary studies. When measuring the performance of subjects, it is important to examine the instrument or task validity. On a related note, including only a small set of test items may raise questions about the potential inferences drawn from the results of the primary studies. Given that, the quantity and quality of instruments used in the primary studies seemed to receive little attention in Norris and Ortega's (2000) synthesis. For example, Cadierno's (1995) study featured an interpretation task consisting of 20 sentences, 10 of which were distractors, and a production task consisting of only five items. Similarly, Hulstijn (1989) administered two retention tests, comprising only four target items in the first test and nine target items in the second test, respectively. In the second test, two items were already used as part of the stimulus set in the learning task. Jourdenais et al., (1995) used only one instrument (i.e., writing task) to show that textual enhancement has an effect on the noticing of target L2 forms. There are other studies Norris and Ortega used that recycle test items for pretest and posttests (e.g., Leow, 1998a) which also raise a question on the validity of the instrument used in the primary studies.

The coding consistency of the primary studies should also be taken into consideration. Researchers should provide evidence of the consistency of their coding, such as inter-rater agreement and intra-rater agreement. In this connection, the coder agreement indexes are measures to ensure consistency and validity with which studies are coded. Though some of the

primary studies have reported a measure, others have failed to satisfy this criterion in Norris and Ortega's (2000) study. For example, Jourdenais et al., (1995) reported no intercoder agreement in coding of think-aloud. Likewise, Alanen (1995) did not provide evidence of inter- or intra-coder agreement on their verbal protocol analysis. Because a number of the primary studies used in Norris and Ortega exhibit such weaknesses, their results should have been omitted or used with caution in Norris and Ortega's final synthesis.

The aforementioned concerns on randomization, instrument validity, and coding consistency can be reduced with more selective procedures for determining which studies should be included for a meta-analysis. In other words, the quality of data samples – and eventually, the quality of meta-analyses – can be improved by applying tighter inclusion criteria or “raising the bar.”

Publication Bias and Inclusion Criteria

The possibility of *publication bias* provides further justification for establishing higher standards for the quality of the primary studies in meta-analysis. Publication bias, a well-known problem in empirical research, refers to a tendency that positive results are published while negative or inconclusive results are not. This is so because authors are more likely to submit manuscripts reporting positive results and less likely to submit (or accept) those with negative or null results (Rosenthal, 1979). Publication bias can harm the validity of meta-analysis. According to Smith (1980), publication bias leads to biased (i.e., overestimated) average effect sizes. This may be the case especially in less prestigious journals. Specifically, the authors of studies with smaller sample size are likely to submit their work to less prestigious journals. Among those studies, studies with negative or null results are likely to be rejected by the editor. As a result, only papers reporting positive effects are likely to be published in those journals. Those studies therein tend to report more positive effects with extreme magnitude due to larger sampling errors. In sum, primary studies from less prestigious journals are likely to have smaller sample size, overestimating average effect size due to publication bias. This argument is consistent with previous research, which indicates that primary studies with small sample sizes tend to report larger effect sizes than studies with large samples (e.g., Rothstein, Sutton, & Borenstein, 2005; Slavin, 2008).

Norris and Ortega (2000) did not use any quality indicators in the selection of their primary studies; instead, they used an “inclusive approach.” In addition, their effect size estimate may also reflect the possibility of overestimation due to publication bias. One way to address such an “inclusive approach” is to conduct a comparative study (or “sensitivity analysis”)⁵ with a set of criteria. The comparative study will determine if the differences between the effect sizes of the two groups (i.e., a stringent sample vs. a less stringent sample, or peer-reviewed vs. non-peer-reviewed) are notable. If there is little difference, we can say the main result is robust. If not, we need further analysis to validate the result before making any judgment call. The comparative study can also employ a journal quality indicator. For example, the Journal Impact Factor from Journal Citation Report (JCR) provides a quantitative tool, if not an objective means, for evaluating the world's leading journals in all disciplines. The Impact Factor measures “the frequency with which the ‘average article’ in a journal has been cited in a particular year or

⁵ The term “sensitivity analysis” was brought to my attention by Dr. Lourdes Ortega (personal communication, March 11, 2010), to whom I am truly indebted.

period” (Garfield, 1994, p. 1).⁶ Even though it would be controversial to determine which journals are considered “more prestigious” than others in the L2 domain, I hope we can reach a scholarly consensus in the near future. After all, a journal quality indicator, once properly set, may alleviate not only the publication bias issue, but also concerns about randomization, instrument validity, and coding consistency. In other words, quality indicators are the key to resolving the data collection issues raised in this section. They improve the quality of data samples, which leads to more reliable average effect sizes and more meaningful meta-analyses. In sum, performing comparative analysis (or sensitivity analysis) based on a set of quality criteria present a viable alternative, and possibly, a useful way to improve the quality of future meta-analysis. The quality of output depends very much on the quality of input.

Coding System Issues

Regarding the coding system used in Norris and Ortega’s (2000) synthesis, the issue of oversimplification is evident in three ways: (1) in its research designs, (2) in its target population and language context, and (3) in the nature of its instructional treatments. I also point out the importance of moderating variables which can be obtained through proper coding schemes.

Experimental vs. Quasi-experimental Research Designs

Depending on the purposes of the primary studies, either experimental or quasi-experimental designs can be employed in a meta-analysis to investigate the relationship among variables of interest. Given that, the use of quasi-experimental research designs in meta-analysis warrants particular attention. While each design has its own advantages and disadvantages, Kerlinger (1986) identifies three major weaknesses of quasi-experimental research design: (1) the inability to manipulate independent variables, (2) the lack of power to randomize, and (3) the risk of improper interpretation.

In their meta-analysis, Norris and Ortega (2000) did not distinguish between the results from true experimental studies and those from quasi-experimental studies. Instead, all the results from the experimental and quasi-experimental studies were pooled together in the computation of the average effect sizes in order to determine the effectiveness of L2 instruction. In fact, 56% of the primary studies in Norris and Ortega followed an experimental research design with control/comparison groups, of which only 17% reported the use of true control groups (Norris & Ortega, 2000). This could be misleading. Given the lack of randomization and proper manipulation of independent variables, which are inherent problems with quasi-experimental design, it would have been desirable to examine the effects from the true experimental studies and those from the quasi-experimental studies separately. By applying a coding scheme that separates experimental and quasi-experimental studies, the effect size statistics from the pooled data could have been compared with those from pure experimental primary studies. This way, one could have seen whether their inclusive approach was robust with respect to research designs.

⁶ Journal Impact Factor has its own weaknesses. For example, the index of one journal may change over time. In addition, the number of citation is only one aspect of journal quality. I am very grateful to Dr. Michael Kieffer (personal communication, April 8, 2010) for raising this issue.

Target Population & Language Setting

Given the assumption that instructional intervention impacts L2 learning in a similar population with similar characteristics, individual characteristics of learners (e.g., L2, L1, proficiency, educational context) should be considered when specifying the target population of interest, i.e., the population to which the researcher(s) aim(s) to generalize the results. In their study, however, Norris and Ortega did not identify the target population of primary studies based on the characteristics of learners: the educational context of the participants included in the synthesis, for example, ranges from elementary to college settings. Even though child SLA and adult SLA can be similar to some degree, research has shown that there is a difference in their rates of development. For instance, Schachter (1988) points out that child second language learners have a greater capacity to reach native-like fluency of the target language more so than adult second language learners, most likely due to the different kinds of knowledge (e.g., their notion of equipotentiality) that these two groups of learners display.

Norris and Ortega (2000) also failed to pay attention to the fact that second language learners are situated in a variety of language contexts in the course of acquiring another language. In the acquisition of English, for example, learners may attend classes in their native countries through which they are taught a “second language” in either their native or foreign language (i.e., English as a foreign language or EFL). Learners may also move to a country where they receive instruction by native speakers of that country (i.e., English as a second language or ESL). As ESL learners interact naturalistically with native speakers, the language exposure or input that is readily available to learners in their respective learning environments are differential, which in turn has an effect on their second language acquisition. Hence, precise specification of the target population and the language contexts of the SLA in the synthesis is necessary for meaningful generalization.

Nature of Instructional Treatments: FonF and FonFS

Over the past 20 years, SLA scholars have expended much effort investigating the relative effectiveness of two pedagogical paradigms, Focus on Form (FonF), and Focus on FormS (FonFS). FonF conceptualizes teaching within a meaningful communication framework in which negotiation for meaning is important, while FonFS entails an extraction of grammatical features that is in line with traditional instruction. In Norris and Ortega’s (2000) study, FonF and FonFS were found to be equally effective in promoting L2 learning. This is rather a startling finding given that a myriad of studies in the L2 literature have suggested that FonF fosters greater outcomes in language learning.

Upon closer examination, this finding could be attributed to the fact that Norris and Ortega (2000) only examined one aspect of FonF. Over the years, however, a number of different definitions have evolved as to what constitutes FonF (e.g., Doughty & Williams, 1998; Long, 1991; Long & Robinson, 1998). That is, FonF instruction can be classified on a continuum bounded by Doughty and Williams’ (1998) definition of preemptive language intervention on one end and Long’s (1991) definition of incidental pedagogical intervention on the other end, depending on the scope of definition. The linguistic elements of focus in Doughty and Williams’ definition are preplanned by the teacher, whereas in Long’s case, they are unplanned. In fact, the instructional treatment can be coded in varying degrees of form-focused instruction. Specifically, a taxonomy of form-focused instruction encompassing FonFS, planned FonF, and incidental

FonF can represent the continuum inherent in FonF and FonFS constructs (Ellis, 2001). Accordingly, modification of the existing coding scheme seems necessary. A case in point would be the fact that the primary studies on input processing (IP) (e.g., VanPatten & Cadierno, 1993) in Norris and Ortega are categorized as FonF explicit instruction. However, a closer analysis of these processing instruction studies reveals that both explicit and implicit treatments are included in the FonF definition. According to VanPatten (1996), processing instruction, the pedagogical intervention of his IP model, has three components: (1) explicit information pertaining to the target form, which pertains to explicit FonF; (2) information not utilized in optimal processing strategies, and (3) structured input activities, in which the orientation is largely concerned with implicit FonF. Thus, a similar argument can be made regarding input enhancement studies. It is highly probable that these studies have been treated by Norris and Ortega as implicit FonF. However, given that perceptual saliency can also be driven internally by the learner (because of readiness, for example), an implicit form of FonF instruction may be experienced in an explicit way by the learner. Because a wide range of characteristics can be coded differently for the same treatment, it is possible that the secondary coding by Norris and Ortega may have led to a biased conclusion.

Coding of Moderating Variables

Despite the fact that moderating variables can be easily obtained with the help of properly developed coding schemes, potential moderating variables have received little attention in L2 meta-analysis. The truth of the matter is: moderating variables are necessary to help explain variance in effect sizes across the primary studies. Moderating variables such as learner aptitude, learner awareness, structural complexity, and frequency of exposure to target-L2 tokens have been found to play a role in instruction effectiveness (Norris & Ortega, 2000). It is also possible that moderating variables such as the proficiency level of participants, their age, language context (e.g., ESL, EFL), target language, and between testing have been useful in explaining the effectiveness of L2 instruction – but all this has not been addressed in Norris and Ortega's study. For example, researchers in the sampled primary studies performed pretests and posttests over different time spans. The amount of time between the tests as well as the duration of the treatment could have affected the results of the primary studies, and thus, the outcome of the synthesis. In other words, time variable can serve as moderating variable. In this light, the time variables should have been controlled in order to estimate a true effect size.

A revealing illustration of how moderating variables can be overlooked is a research synthesis study by Lee and Huang (2008), which follows the methodological convention of Norris and Ortega (2000). The main goal of the synthesis was to examine the impact of visual input enhancement on grammar learning. In their study, 16 primary studies contributed 20 unique samples for the research synthesis. Results of the study indicated that a small effect size ($d = 0.22$) was obtained for learners who were exposed to enhanced texts (Appendix A). The effect size might be misleading, however, because Lee and Huang compiled all data samples with different target languages. When one calculates the average effect sizes for Lee and Huang's data samples using target languages (i.e., English and Spanish) as a moderating variable, the outcomes became considerably different from the results found and reported by Lee and Huang. Using the effect size values reported in their synthesis, each primary study was matched with its respective target language, and new effect size statistics were computed as shown in Table 1. The mean effect sizes for English target language studies were statistically significant ($d = 0.37$) while mean effects for Spanish target language studies were not

($d = -0.01$). This re-calculation shows target language is a potential and possibly important moderating variable. This example shows that the potential effect of moderators may be critical in exploring the treatment effectiveness through meta-analysis.

TABLE 1
A re-calculation of the effect size in Lee and Huang's (2008) study

Study	Effect size (d)				
	Grammar Learning				Meaning comprehension
	Target Language	Immediate posttest	Delayed posttest	Pre-to-post contrast	
Doughty (1991)	English	0.46	n.t.	1.85	n.a.
Ha (2005)	English	0.07	n.t.	0.26	n.t.
Izumi (2002)	English	0.02	n.t.	0.67	n.t.
Lee (2007)	English	1.05	n.t.	1.11	-0.73
White (1998)	English	0.26	-0.01	0.97	n.t.
Mean		0.37	-0.01	0.97	-0.73
<i>SD</i>		0.42	n.a.	0.59	n.a.
<i>K</i>		5	1	5	1
95%CI upper		0.74		1.49	
95%CI lower		0.01		0.46	
Study	Target Language	Immediate posttest	Delayed posttest	Pre-to-post contrast	Meaning comprehension
Jourdenais 1 (1998)	Spanish	-0.04	-0.1	-0.04	n.a.
Jourdenais 2 (1998)	Spanish	-0.02	-0.1	-0.19	n.a.
Jourdenais 3 (1998)	Spanish	-0.15	0.09	0.01	n.a.
Kubota 1 (2000)	Spanish	-0.37	-0.45	n.a.	n.t.
Kubota 2 (2000)	Spanish	-0.1	-0.21	n.a.	n.t.
Leow (1997)	Spanish	-0.06	n.t.	0.86	0.28
Leow (2001)	Spanish	n.a.	n.t.	n.a.	n.a.
Leow et al (2003)	Spanish	n.a.	n.t.	n.a.	n.a.
Overstreet 1 (1998)	Spanish	0.07	n.t.	0.07	-0.94
Overstreet 2 (2002)	Spanish	0.61	n.t.	n.t.	-0.34
Overstreet 1 (2002)	Spanish	-0.05	n.t.	n.t.	-0.13
Shook (2004)	Spanish	n.a.	n.t.	n.a.	-0.30
Mean		-0.01	-0.15	0.14	-0.29
<i>SD</i>		0.26	0.20	0.41	0.44
<i>K</i>		9	5	5	5
95%CI upper		0.16	0.02	0.50	0.10
95%CI lower		-0.18	-0.33	-0.22	-0.67

Note, n.t. = not tested in the particular study; n.a. = not applicable because the study did not provide sufficient data for the effect size calculations; k = number of samples contributing to the effect size calculation.⁷

⁷ Lee and Huang's (2008) notations are used here.

Statistical Analysis Issues

One of the key challenges in implementing a meta-analysis is acquiring an effect size through statistical analysis procedures. First, decisions need to be made among different effect size statistics in order to compute the effect size of an individual primary study. Then, statistical approaches would need to be carefully selected for the purpose of summarizing the individual effect sizes from the primary studies. According to Lipsey and Wilson (2001), this decision rests on the “nature of the research findings, the statistical forms in which they are reported, and the hypotheses being tested by the meta-analysis” (p. 34). In other words, both the characteristics of the primary studies as well as the objectives of the synthesis should be taken into account. In the following section, two main statistical approaches to calculating and summarizing the effect sizes are presented. By describing the different properties of the effect size statistics and analyzing the characteristics of the primary studies, I suggest that future researchers consider statistical approaches other than the one used by Norris and Ortega (2000).

Computing Individual Effect Size: Cohen’s d vs. Hedges’ g

Meta-analysts have used diverse effect size indices to measure the magnitude of effect. According to Cohen (1977), effect size can be defined as “the degree to which the phenomenon exists” (p. 4). The theoretical assumptions of effect size indices and their corresponding formulas can be rather complex, and thus, I omit specific explanations. Instead, in this section, I briefly summarize the two most commonly used effect size statistics in the meta-analysis literature, namely Cohen’s d and Hedges’s g . Meta-analysts should reflect on their appropriate usage, which depends on data characteristics.

According to Lipsey and Wilson (2001), the unstandardized effect size index is used when the pre-post test findings to be meta-analyzed entail the same operationalization of the variables (e.g., the same measure) in the research synthesis. In this case, the mean difference scores from different samples are comparable, and thus, the average effect size is obtained using the formula (1.0).

$$(1.0) \quad ES = \bar{X}_1 - \bar{X}_2$$

However, pre-post test findings often tend to entail different operationalizations across studies, implying that the unstandardized effect size index is of little use in practice. In this case, a standardized method such as Cohen’s d or Hedges’ g must be used. The standardized approach divides effect size by a pooled standard deviation to control for heterogeneity between pre- and post-tests. Cohen’s d is defined as the difference between the two means divided by a pooled standard deviation as defined in (2.1). Note that Cohen (1977) originally assumed that standard deviations of pre- vs. post-test groups are equal. Later researchers have often modified the formula for pooled standard deviation, resulting in various versions of Cohen’s d .

$$(2.1) \quad \text{Cohen's } d: d = \frac{\bar{X}_1 - \bar{X}_2}{S_p}, \quad S_p = S_1 = S_2$$

Hedges’ g is defined like the Cohen’s d with the exception of how a pooled standard deviation is estimated, as in (2.2).

$$(2.2) \text{ Hedges' } g: g = \frac{\bar{X}_1 - \bar{X}_2}{s_p}, \quad s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)}}$$

Sampling Error Difference Problem in Summarizing Individual Effect Sizes

After obtaining the Cohen's *d* or Hedges' *g* of the primary studies, meta-analysts should report the summary statistic of the individual effect sizes. In general, researchers compute unweighted or weighted average effect size using the individual effect sizes of primary studies. In so doing, dealing with sampling error difference problems poses a challenge.

Accounting for differences in sampling error between primary studies with differing sample sizes is regarded as an extremely important consideration in meta-analysis (Lipsey & Wilson, 2001). Sampling error, "the error that occurs when a statistic based on a sample estimates or predicts the value of a population parameter" (Agresti & Finlay, 1997, p. 23), is greater for smaller samples than for larger samples. In the meta-analysis context, each effect size from a primary study represents one data sample taken from a given population. It follows that each one of those data samples is likely to be different from the true effect size by some error component or the sampling error. Therefore, the impact of an idiosyncratic sampling error due to different sample size across primary studies should be seriously considered when the average size effect is computed.

What can one make of the statistical procedure of summarizing effect sizes adopted by Norris and Ortega (2000)? In short, the researchers obtained the average effect size by simply aggregating individual effect sizes of the 49 data samples from 45 primary studies and then by dividing it with 49. In other words, the 49 data samples were given equal weight when computing the average effect size.⁸ In so doing, the possible impact of idiosyncratic sampling errors across the primary studies was not taken into consideration. The consequence of this oversight can be observed in Figures 1-4, of which only the original figure (i.e., "Figure 2. Effect sizes plotted against study group samples size for 78 unique sample studies") is reproduced from Norris and Ortega. The issue of sampling error difference is unraveled through my illustration of this one figure; it is helpful to view the summary visually. The x-axis represents effect size statistics and the y-axis represents group sample size. The majority of the primary studies found in Figure 1 are based on a very small sample (e.g., $N < 10$), while a few studies employ relatively large numbers of subjects (e.g., $N > 30$). Studies with larger sample sizes (e.g., greater than 30) report a tighter effect size distribution, ranging from 0.70 to 2.50 as shown in Figure 1. On the other hand, studies with small sample sizes (e.g., $N < 30$) exhibit a wide dispersion of effect sizes, ranging from -1.5 to 3.5 (Figure 1). This shows that the variability in effect size is larger for the primary studies with smaller sample sizes than for the primary studies with larger sample sizes. In other words, a smaller sample size may lead to a bigger sampling error, and thus, ignoring the difference in sample size across the studies may lead to a biased average effect size.

It becomes clearer that the effect sizes were in fact skewed to the right once the axis on the graph is centered on zero (Figure 2). Indeed, Norris and Ortega (2000) attempted to avoid this problem by saying small size studies might be statistically insignificant. But this does not explain why the effect sizes are skewed to the right. In fairness, studies with effect size greater than +1.5 are observed in Figure 3. These studies exhibit a somewhat extreme effect size

⁸ Even though Norris and Ortega computed 95% CI for the average effect, this approach still does not consider heterogeneity with respect to sampling error due to different sample sizes across the primary studies.

magnitude which could potentially exaggerate the average effect size, which is consistent with the publication bias issue as explained earlier. Figure 3 further implies that if the sampling error had been taken into account in the analysis, meaning that more weight was given to the studies with larger sample size and less weight to studies with smaller sample size, the effect size estimate might have shifted to the left.

According to Cohen (1977, 1988), an effect size of .8 or higher is considered large, an effect size of .5 through .8 indicates a moderate effect and an effect size of .5 or below is considered a small effect. The effect size reported by Norris and Ortega (2000) for instructional effectiveness is 0.96. This seems quite large, considering that an effect size of 0.80 and above reflects a high impact on the intervention group. I suspect that this effect size could even be misleading since idiosyncratic sampling errors across the primary studies are ignored in their synthesis. Thus, one would expect the effect size to shrink, i.e., to shift towards the left as illustrated by Figure 4, once the sampling error difference problem is properly addressed.

FIGURE 1

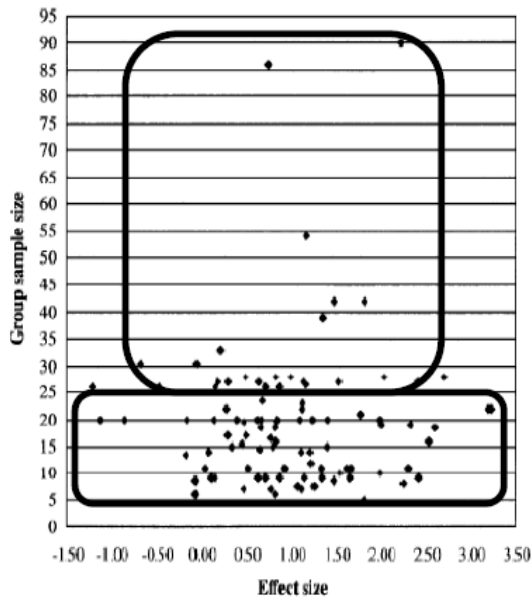


Figure 2. Effect sizes plotted against study group sample sizes for 78 unique sample studies (mean effect size, $d = 0.96$)

FIGURE 2

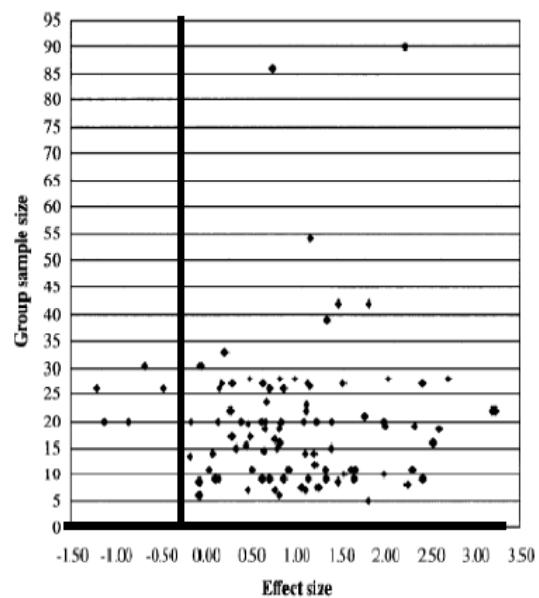


Figure 2. Effect sizes plotted against study group sample sizes for 78 unique sample studies (mean effect size, $d = 0.96$)

FIGURE 3

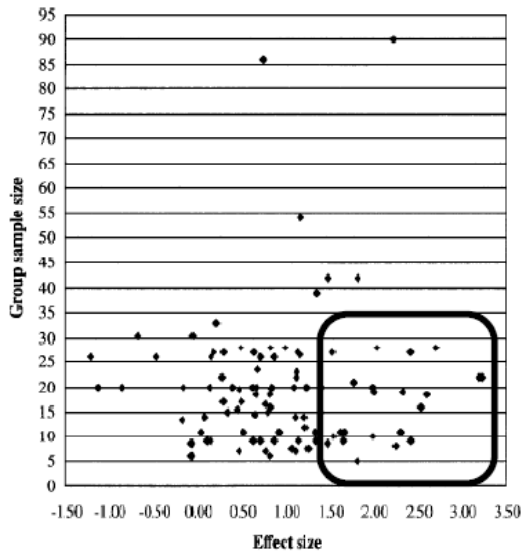


Figure 2. Effect sizes plotted against study group sample sizes for 78 unique sample studies (mean effect size, $d = 0.96$)

FIGURE 4

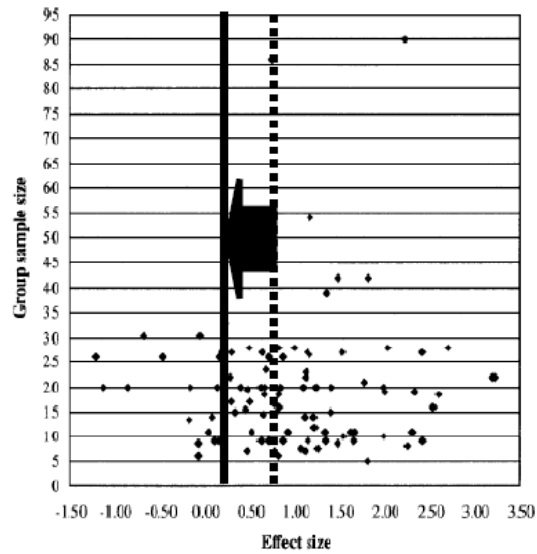


Figure 2. Effect sizes plotted against study group sample sizes for 78 unique sample studies (mean effect size, $d = 0.96$)

To deal with this problem, one could re-compute and compare the average effect sizes based on the sample size (e.g., $N = (n_1 + n_2) > 20$, $N > 30$, $N > 50$). By this supplementary procedure, one might obtain a better understanding of the efficacy of instruction. In sum, Norris and Ortega's (2000) estimated effect size is subject to diverse sources of variation, namely, publication bias and the sampling error difference problem, issues that *can* be accounted for with rigorous statistical approaches.

Correcting the Sampling Error Differences: Hedges' adjusted g

The figures above demonstrate the need for meta-analysts to re-examine the primary studies with extreme effect size statistics. Specifically, are the high effect sizes obtained from the primary studies published in non-peer reviewed journals and/or from less prestigious journals? It calls for further investigation of these studies to see if there is indeed an underlying systematic problem such as publication bias. This issue would be important to achieve external validity in a meta-analysis, where the data can potentially influence the outcomes. While publication bias cannot be completely removed; it can be minimized by using a proper quality index.

Assuming that a proper quality index is employed in the future L2 meta-analyses, the more practical question is: Would it be possible to incorporate small sample size studies, accounting for the upward bias of the population effect size? In fact, Hedges (1981) pointed out that Hedges' g can lead to an upward bias in effect size for small samples ($N = n_1 + n_2$), specifically when N is smaller or equal to 20.⁹ To correct the upward bias, Hedges suggested a small sample bias correction procedure, which is known as Hedges' adjusted g . As shown in the formula below (2.3), a small sample bias can be corrected by simply multiplying a factor of k .

⁹ Some textbooks quote 10.

(2.3) Hedges' adjusted g : $g' = k * g$, $k = \left[1 - \frac{3}{4(n_1 + n_2) - 9} \right]$

Note that as $n_1 + n_2$ becomes larger, k converges to 1. However, when $n_1 + n_2$ is relatively small, the effect size from a study will be corrected up to factor k . In other words, Hedges' adjusted g gives more weight to the primary studies with larger sample sizes in computing the average effect size across studies. This weighted means increases the importance of large sample studies among which small sample sizes studies tend to exaggerate effect sizes (Rothstein, Sutton, & Borenstein, 2005; Slavin, 2008). Given the data characteristics of Norris and Ortega (2000), the study would have benefited from the use of Hedge's adjusted g approach, in which the mean effect sizes are weighted by sample size in pooling the effect sizes across the primary studies. The table below summarizes the effect size types discussed thus far and their relationship to the nature of data.

TABLE 2
Summary of the effect size statistics

Effect Size Type	Two-Variable Relationships	
	$S_1 \neq S_2$	$n_1 + n_2$ small
Unstandardized	No	No
Cohen's d	No	No
Hedges' g	Yes	No
Hedges' adjusted g	Yes	Yes

Prevalent Sampling Error Difference Problems in L2 Meta-Analysis

Sampling error differences seem to be particularly important for L2 primary studies because these studies tend to have small sample sizes. Without accounting for the heterogeneity across the primary studies, findings will not be as meaningful and insightful regardless of the number of the data synthesized. One of the most important issues for meta-analysts in applied linguistics and TESOL is to correct this inconsistency.

Norris and Ortega (2000) explicitly state that there is an overall pattern of sampling errors when comparing group sample sizes of 20 or more and group sample sizes of 20 or fewer. However, they do not state how they chose the baseline sample group size of 20 to investigate such a pattern observed possibly due to sampling error. This, in turn, became a subjective criterion, since different baselines (e.g., 30, 50, and even 100) may produce different patterns. Even more importantly, Norris and Ortega do not offer a solution to the problem they point out; they only illustrate the effect sizes for two groups (i.e., $N \geq 20$ vs. $N < 20$).

Since the introduction of Norris and Ortega's (2000) research synthesis, subsequent L2 meta-analysts have relied on their synthesis mode, but in these past 10 years, researchers have made little advances in addressing the methodological weaknesses in Norris and Ortega's synthesis. Given that Hedges' adjusted g has been available for the past 30 years, for example, it is somewhat striking to see that no attempt has been made to account for idiosyncratic sampling errors across data samples by subsequent L2 meta-analysts. A critical analysis of Norris and Ortega's methodological procedures is imperative. Table 3 displays a list of research syntheses which followed the methodological procedures of Norris and Ortega. One synthesis that did take precautions in estimating effect size is Mackey and Goo (2007). Reflecting on the need to

examine the nature of data, Mackey and Goo “obtained corrected (or unbiased) effect size estimates based on the sample sizes and used the inverse variance weight method to get weighted mean effect sizes, a commonly accepted method for weighting in meta-analysis” (p. 419). Considering the gravity of the idiosyncratic sampling error issue in Norris and Ortega’s synthesis, one could raise concerns about the validity of previous research syntheses conducted between 2000 and 2008¹⁰ in SLA, with the exception of Mackey and Goo’s synthesis.

TABLE 3
Summary of subsequent L2 meta-analyses

L2 meta-analysis
“The ‘ <i>d</i> -index’ (the effect size that is usually associated with t-tests or F-tests based on a comparison of two treatment levels) was calculated for each of the studies under review. It was calculated using the following formula: $d = \frac{X_1 - X_2}{SD_1 + SD_2 / 2}$ where X_1 and X_2 = the two group means; and SD_1 and SD_2 = the pooled standard deviation for the two groups.” (Cooper, 1998, as cited in Dinsmore, 2006, p. 69)
“The results from the 13 collected research studies were aggregated by calculating Cohen’s <i>d</i> (effect size), which can be interpreted as the magnitude of an observed difference between two groups in standard deviation units (Norris & Ortega 2000) ... We followed the procedure for calculating Cohen’s <i>d</i> that is explained in detail by Lipsey and Wilson (2001).” (Jeon & Kaya, 2006, p. 177)
“To compare the effect of treatment against control/comparison groups, as well as group change between pretests and posttest, we used Cohen’s <i>d</i> (adapted from Norris & Ortega, 2000, p. 442).” (Keck, Iberri-Shea, Tracy-Ventura, & Wa-Mbaleka, 2006, pp. 105-106)
“The next step in the process was to calculate effect sizes for the 31 studies. To do this, we used Wilson’s (2001) <i>Effect Size Determination Program</i> to calculate Cohen’s <i>d</i> values.” (Russell & Spada, 2006, p. 146)
“I will rely on the measure most widely used, Cohen’s <i>d</i> , which is the number of standard deviations by which the means of two groups differ.” (Truscott, 2007, p. 256)
“We followed Norris and Ortega’s (2000) formulas for the calculation of effect sizes and confidence intervals.” (Lee & Huang, 2008, p. 327)

Alternative Procedure to Meta-Analysis: Hierarchical Linear Model (HLM)

Note that Hedges’ adjusted *g* is only an approximate way of correcting the upward bias due to idiosyncratic sampling errors. The hierarchical linear model, or HLM (Raudenbush & Bryk, 2002), is a more systematic and statistically rigorous approach to addressing the sampling error difference issue. Since the data in a meta-analysis are equivalent to descriptive summary statistics of the primary research studies, and not raw data, the participants of a study are nested within the sample of primary studies. HLM uses a hierarchical structure, in which subjects are

¹⁰ The L2 meta-analyses reviewed here are those studies conducted through 2008.

nested within the primary studies included in the research synthesis (Raudenbush & Bryk, 2002). This framework enables the meta-analyst to explicitly account for both sampling error and inconsistency in the effect size of the corresponding parameter. HLM can be particularly helpful in correcting a bias that occurs when the sample size is very small or very large (Hedges, 1981). By using precision-weighted estimates, we can separate the sources of error attributed to the heterogeneity of effect sizes across the primary studies. Furthermore, the use of HLM is known to yield more accurate estimates on the parameter of interest (e.g., the effectiveness of L2 instruction) by taking into account the correlations among sampled studies within a study (within-study level-one model) and at the same time by controlling for key covariates such as time, or interaction between the target language and L2 instruction (between-studies level-two model). Thus, HLM would not only correct some bias but also provide a more meaningful comparison across the primary studies. Another advantage of HLM is that the framework allows us to systematically incorporate moderating variables in the model. This aspect is particularly useful, considering the unique characteristics of L2 data samples as discussed in the previous section. With all these strengths in its favor, HLM may become a useful statistical vehicle for future meta-analyses in applied linguistics.

CONCLUSION

The impact of Norris and Ortega's (2000) study on second language research has been exemplary: it has filled gaps in second language research and provided new directions for L2 researchers in the past decade. In the nine years since Norris and Ortega's publication, many researchers have adopted their methodological procedures with insightful results.¹¹ Nonetheless, the time has come to revisit and reexamine their methodological procedures and the associated implications for future L2 meta-analyses. I hope that the concerns raised here give an overview both of critical issues and of possible directions for future research.

Notwithstanding their notable contributions, Norris and Ortega (2000) made decisions that subsequent researchers have followed uncritically. When seven subsequent L2 meta-analyses were examined, some of the issues that appeared in the Norris and Ortega's meta-analysis were perpetuated. All seven studies adopted the "inclusive approach," paying little attention on the data quality issue (Table 4). In addition, six out of seven studies employed the Cohen's *d* index, ignoring the problem of sampling error across primary studies.

¹¹ This paper first began as a class assignment in fall 2008, which I later developed into my first year doctoral student paper in spring 2009. At the time this critical study was completed, it had been nine years to be exact.

TABLE 4
Influence of Norris and Ortega (2000) on subsequent L2 meta-analysis

	L2 meta-analysis (N=7)	L2 meta-analysis
Inclusive approach	7/7	Dinsmore (2006) Jeon & Kaya (2006) Keck, Iberri-Shea, Tracy-Ventura, & Wa-Mbaleka (2006)
Cohen's <i>d</i>	6/7	Lee & Huang (2008) Mackey & Goo (2007) Russell & Spada (2006) Truscott (2007)

Given the methodological issues of the Norris and Ortega (2000) study, conclusions drawn from their synthesis need to be carefully interpreted. Norris and Ortega themselves acknowledged some “caveats” of their meta-analysis, cautioning readers of the limitations of their findings. As reaffirmed by Han (2004), “It is imperative to interpret the findings from the Norris and Ortega (2000) study as suggestive rather than definitive” (p. 129). In light of the current review, I propose qualitative critiques on the same body of quantitative research findings.¹² Slavin (1986) calls this approach to meta-analysis “best evidence synthesis,” as conclusions are drawn from both quantitative and qualitative reviews. A mixed-methods approach can draw more insightful information from syntheses. A more rigorous methodology, such as HLM, coupled with more stringent inclusion criteria, would likely yield more meaningful and useful results for researchers and practitioners in applied linguistics.

ACKNOWLEDGMENTS

I am grateful to Dr. ZhaoHong Han and the two anonymous reviewers for their invaluable suggestions. I am also grateful to Dr. Lourdes Ortega and Dr. Michael Kieffer for their insights and comments. Lastly, I would like to thank Dr. Doug Flahive, Adrienne Wai Man Lew, Tim Hall, and my cohort in the spring 2009 doctoral seminar for their comments on an earlier version of this paper.

REFERENCES¹³

Agresti, A., & Finlay, B. (1997). *Statistical methods for the social sciences* (3rd ed.). New Jersey: Prentice Hall.

* Alanen, R. (1995). Input enhancement and rule presentation in second language acquisition. In

¹² I am grateful to Dr. ZhaoHong Han for raising this issue in our earlier discussion on meta-analysis (October 2, 2008).

¹³ One asterisk (*) indicates those primary studies included in Norris and Ortega (2000) study to calculate an effect size estimate.

- R. Schmidt (Ed.), *Attention and awareness in foreign language learning and teaching* (Technical Report No. 9) (pp. 259-302). Honolulu, HI: University of Hawai'i, Second Language Teaching & Curriculum Center.
- * Bouton, L. F. (1994). Can NNS skill in interpreting implicature in American English be improved through explicit instruction?-A pilot study. *Pragmatics and Language Learning*, 5, 89-109. (ERIC ED 398742).
- * Cadierno, T. (1995). Formal instruction from a processing perspective: An investigation into the Spanish past tense. *The Modern Language Journal*, 79, 179-193.
- * Carroll, S., & Swain, M. (1993). Explicit and implicit negative feedback: An empirical study of the learning of linguistic generalizations. *Studies in Second Language Acquisition*, 15, 357-386.
- * Carroll, S., Roberge, Y., & Swain, M. (1992). The role of feedback in adult second language acquisition: Error correction and morphological generalization. *Applied Psycholinguistics*, 13, 173-189.
- Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (Rev. ed.). New York: Academic Press.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Cooper, H., & Hedges, L. V. (1994). Research synthesis as a scientific enterprise. In H. Cooper, & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 3-14). Thousand Oaks, CA: Sage.
- Cooper, H. M. (1982). Scientific guidelines for conducting integrative research reviews. *Review of Educational Research*, 52, 291-301.
- * Day, E., & Shapson, S. (1991). Integrating formal and functional approaches to language teaching in French immersion: An experimental study. *Language Learning*, 41, 25-58.
- * de Graaff, R. (1997). The eXperanto experiment: Effects of explicit instruction on second language acquisition. *Studies in Second Language Acquisition*, 19, 249-297.
- * DeKeyser, R. M. (1995). Learning second language grammar rules: An experiment with a miniature linguistic system. *Studies in Second Language Acquisition*, 17, 379-410.
- * DeKeyser, R. M. (1997). Beyond explicit rule learning: Automating second language morphosyntax. *Studies in Second Language Acquisition*, 19, 195-221.
- * DeKeyser, R. M., & Sokalski, K. J. (1996). The differential role of comprehension and production practice. *Language Learning*, 46, 613-642. [Two studies].
- Dinsmore, T. H. (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* (pp. 53-90). Amsterdam: John Benjamins.
- Dörnyei, Z. (2007). *Research methods in applied linguistics: Quantitative, qualitative and mixed methodologies*. Oxford: Oxford University Press.
- * Doughty, C. (1991). Second language instruction does make a difference: Evidence from an empirical study of SL relativization. *Studies in Second Language Acquisition*, 13, 431-469.
- Doughty, C., & Williams, J. (1998). Pedagogical choices in focus on form. In C. Doughty & J. Williams (Eds.), *Focus on form in classroom second language acquisition* (pp. 197-261). New York: Cambridge University Press.
- Ellis, R. (2001). Introduction: Investigating form-focused instruction. *Language Learning*, 51, 1-46.

- * Ellis, R., Rosszell, H., & Takashima, H. (1994). Down the garden path: Another look at negative feedback. *JALT Journal*, 16, 9-24.
- * Fotos, S., & Ellis, R. (1991). Communicating about grammar: A task-based approach. *TESOL Quarterly*, 25, 605-628. [Two studies].
- Garfield, E. (1994). The Thomson Reuters Impact Factor. Retrieved October 7, 2008 from http://thomsonreuters.com/products_services/science/free/essays/impact_factor/
- Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher*, 5, 3-8.
- Green, B., & Hall, J. (1984). Quantitative methods for literature review. *Annual Review of Psychology*, 35, 37-53.
- Han, Z.-H. (2004). *Fossilization in adult second language acquisition*. Clevedon: Multilingual Matters.
- * Harley, B. (1989). Functional grammar in French immersion: A classroom experiment. *Applied Linguistics*, 10, 331-359.
- Hedges, L.V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of the American Statistical Association*, 74, 311-319.
- * Herron, C., & Tomasello, M. (1988). Learning grammatical structures in foreign language: Modelling versus feedback. *The French Review*, 61, 910-922.
- * Hulstijn, J. H. (1989). Implicit and incidental second language learning: Experiments in the processing of natural and partly artificial input. In H. W. Dechert & M. Raupach (Eds.), *Interlingual processes* (pp. 49-73). Tübingen: Gunter Narr. [Two studies].
- Hunter, J., Schmidt, F., & Jackson, G. (1982). *Meta-analysis: Cumulating research findings across studies*. Beverly Hills: Sage Publications.
- Jeon, E. H., & Kaya, T. (2006). Effects of L2 instruction on interlanguage pragmatic development: A meta-analysis. In J. M. Norris & L. Ortega (Eds.), *Synthesizing research on language learning and teaching* (pp. 165-211). Amsterdam: John Benjamins.
- * Jourdenais, R., Ota, M., Stauffer, S., Boyson, B., & Doughty, C. (1995). Does textual enhancement promote noticing? A think-aloud protocol analysis. In R. Schmidt (Ed.), *Attention and awareness in foreign language learning* (Technical Report No. 9) (pp. 183-216). Honolulu, HI: University of Hawai'i, Second Language Teaching & Curriculum Center.
- Keck, C. M., Iberri-Shea, G., Tracy-Ventura, N., & Wa-Mbaleka, S. (2006). Investigating the empirical link between task-based interaction and acquisition: A meta-analysis. In J. M. Norris & L. Ortega (Eds.), *Synthesizing research on language learning and teaching* (pp. 91-131). Amsterdam: John Benjamins.
- Kerlinger, R. (1986). *Foundations of behavioral research* (3rd ed.). New York: CBS College Publishing.
- * Kubota, M. (1994). The role of negative feedback on the acquisition of the English dative alternation by Japanese college students of EFL. *Institute for Research in Language Teaching Bulletin*, 8, 1-36. (ERIC ED 386023).
- * Kubota, M. (1995a). The garden path technique: Is it really effective? *Working Papers of Chofu Gakuen Women's Junior College*, 27, 21-48. (ERIC ED 386021).
- * Kubota, M. (1995b). Teachability of conversational implicature to Japanese EFL learners. *Institute for Research in Language Teaching Bulletin*, 9, 35-67. (ERIC ED 397640).
- * Kubota, M. (1996). The effects of instruction plus feedback on Japanese university students of

- EFL: A pilot study. *Bulletin of Chofu Gakuen Women's Junior College*, 18, 59-95. (ERIC ED 397641).
- Lee, S.-K., & Huang, H. T. (2008). Visual input enhancement and grammar learning: A meta-analytic review. *Studies in Second Language Acquisition*, 30, 307-330.
- * Leow, R. P. (1998a). The effects of amount and type of exposure on adult learners' L2 development in SLA. *The Modern Language Journal*, 82, 49-68.
- * Leow, R. P. (1998b). Toward operationalizing the process of attention in SLA: Evidence for Tomlin and Villa's (1994) fine-grained analysis of attention. *Applied Psycholinguistics*, 19, 133-159.
- * Leow, R. P. (1997). Attention, awareness, and foreign language behavior. *Language Learning*, 47, 467-506.
- Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.
- Long, M. (1991). Focus on form: A design feature in language teaching methodology. In K. deBot, R. Ginsberg, & C. Kramsch (Eds.), *Foreign language research in cross-cultural perspective* (pp. 39-52). Amsterdam: John Benjamins.
- Long, M. (1983). Does second language instruction make a difference? A review of the research. *TESOL Quarterly*, 17, 357-382.
- * Long, M. H., Inagaki, S., & Ortega, L. (1998). The role of implicit negative feedback in SLA: Models and recasts in Japanese and Spanish. *The Modern Language Journal*, 82, 357-371. [Two studies].
- Long, M., & Robinson, P. (1998). Focus on form: Theory, research, and practice. In C. Doughty & J. Willimas (Eds.), *Focus on form in classroom second language acquisition* (pp.15-41). New York: Cambridge University Press.
- * Loschky, L. (1994). Comprehensible input and second language acquisition: What is the relationship? *Studies in Second Language Acquisition*, 16, 303-323.
- * Lyster, R. (1994). The effect of functional-analytic teaching on aspects of French immersion students' sociolinguistic competence. *Applied Linguistics*, 15, 263-287.
- Mackey, A., & Goo, J. M. (2007). Interaction research in SLA: A meta-analysis and research synthesis. In A. Mackey (Ed.), *Input, interaction and corrective feedback in L2 learning* (pp. 407-452). New York: Oxford University Press.
- * Mackey, A., & Philp, J. (1998). Conversational interaction and second language development: Recasts, responses, and red herrings? *The Modern Language Journal*, 82, 338-356.
- * Master, P (1994). The effect of systematic instruction on learning the English article system. In T. Odlin (Ed.), *Perspectives on pedagogical grammar* (pp. 229-252). Cambridge: Cambridge University Press.
- * Nagata, N. (1993). Intelligent computer feedback for second language instruction. *The Modern Language Journal*, 77, 330-339. [Same study as Nagata & Swisher, 1995].
- * Nagata, N. (1995). An effective application of natural language processing in second language instruction. *CALICO Journal*, 13, 47-67.
- * Nagata, N. (1997a). The effectiveness of computer-assisted metalinguistic instruction: A case study in Japanese. *Foreign Language Annals*, 30, 187-200.
- * Nagata, N. (1997b). An experimental comparison of deductive and inductive feedback generated by a simple parser. *System*, 25, 515-534.
- * Nagata, N. (1998). Input vs. output practice in educational software for second language acquisition. *Language Learning & Technology*, 1(2), 23-40.
- Norris, J., & Ortega, L. (2000). Effectiveness of L2 instruction: A research synthesis and

- quantitative meta-analysis. *Language Learning*, 50, 417-528.
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Newbury Park, CA: Sage Publications.
- * Robinson, P. (1996a). *Consciousness, rules, and instructed second language acquisition*. New York: Peter Lang. [Same study as Robinson, 1996b].
- * Robinson, P. (1997). Generalizability and automaticity of second language learning under implicit, incidental, enhanced, and instructed conditions. *Studies in Second Language Acquisition*, 19, 233-247.
- Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, 86(3), 638-641.
- Rothstein, H. R., Sutton, A. J., & Borenstein, M. (Eds.) (2005). *Publication bias in meta-analysis: Prevention, assessment and adjustments*. Chichester, West Sussex, England: John Wiley & Sons.
- Russell, J., & Spada, N. (2006). The effectiveness of corrective feedback for the acquisition of L2 grammar: A meta-analysis of the research. In J. M. Norris & L. Ortega (Eds.), *Synthesizing research on language learning and teaching* (pp. 133-164). Amsterdam: John Benjamins.
- * Salaberry, M. R. (1997). The role of input and output practice in second language acquisition. *The Canadian Modern Language Review*, 53, 422-451.
- Schachter, J. (1988). Second language acquisition and its relationship to Universal Grammar. *Applied Linguistics*, 9, 219 - 235.
- * Scott, V. (1989). An empirical study of explicit and implicit teaching strategies in French. *The Modern Language Journal*, 72, 14-22.
- * Scott, V. M. (1990). Explicit and implicit grammar teaching: New empirical data. *The French Review*, 63, 779-788.
- Slavin, R. E. (2008). What works? Issues in synthesizing education program evaluations. *Educational Researcher*, 37(1), 5-14.
- Slavin, R.E. (1986). Best-evidence synthesis: An alternative to meta-analytic and tradition reviews. *Educational Researcher*, 15, 5-11.
- Smith, M. L. (1980). Publication bias in meta-analysis. *Evaluation in Education: An International Review Series*, 4, 22-24.
- Truscott, J. (2007). The effect of error correction on learners' ability to write accurately. *Journal of Second Language Writing*, 16, 255-272.
- * van Baalen, T. (1983). Giving learners rules: A study into the effect of grammatical instruction with varying degrees of explicitness. *Interlanguage Studies Bulletin Utrecht*, 7, 71-100.
- VanPatten, B. (1996). *Input processing and grammar instruction: Theory and research*. Norwood, NJ: Ablex Publishing.
- * VanPatten, B., & Cadierno, T. (1993). Explicit instruction and input processing. *Studies in Second Language Acquisition*, 15, 225-241.
- * VanPatten, B., & Oikkenon, S. (1996). Explanation versus structured input in processing instruction. *Studies in Second Language Acquisition*, 18, 495-510.
- * VanPatten, B., & Sanz, C. (1995). From input to output: Processing instruction and communicative tasks. In F. Eckman, D. Highland, P. Lee, J. Mileham, & R. Weber (Eds.), *SLA theory and pedagogy* (pp. 169-185). Hillsdale, NJ: Lawrence Erlbaum.
- * White, L., Spada, N., Lightbown, P., & Ranta, L. (1991). Input enhancement and L2 question

formation. *Applied Linguistics*, 12,416-432. [Same study sample as Spada & Lightbown, 1993].

- * Williams, J., & Evans, J. (1998). What kind of focus and on which forms? In C. Doughty & J. Williams (Eds.), *Focus on form in classroom second language acquisition* (pp. 139-155). Cambridge: Cambridge University Press.
- * Yang, L., & Givón, T. (1997). Benefits and drawbacks of controlled laboratory studies of second language acquisition. *Studies in Second Language Acquisition*, 19, 173-194.

APPENDIX A

Summary of effect size *d* values of primary studies in Lee & Huang (2007)

Study	Effect size (<i>d</i>)			
	Grammar Learning			Meaning Comprehension
	Immediate posttest	Delayed posttest	Pre-to-post contrast	
Alanen (1995)	0.08	n.t.	n.t.	n.t.
Doughty (1991)	0.46	n.t.	1.85	n.a.
Ha (2005)	0.07	n.t.	0.26	n.t.
Izumi (2002)	0.02	n.t.	0.67	n.t.
Jourdenais (1998) 1	-0.04	-0.10	-0.04	n.a.
Jourdenais (1998) 2	-0.02	-0.10	-0.19	n.a.
Jourdenais (1998) 3	-0.15	0.09	0.01	n.a.
Jourdenais et al. (1998)	1.61	n.t.	n.t.	n.t.
Kubota (2000) 1	-0.37	-0.45	n.a. ^a	n.t.
Kubota (2000) 2	-0.1	-0.21	n.a.	n.t.
Lee (2007)	1.05	n.t.	1.11	-0.73
Leow (1997)	-0.06	n.t.	0.86	0.28
Leow (2001)	n.a.	n.t.	n.a.	n.a.
Leow et al. (2003)	n.a.	n.t.	n.a.	n.a.
Overstreet (1998)	0.07	n.t.	0.07	-0.94
Overstreet (2002) 1	0.61	n.t.	n.t.	-0.34
Overstreet (2002) 2	-0.05	n.t.	n.t.	-0.13
Shook (1994)	n.a.	n.t.	n.a.	-0.30 ^b
White (1998) ^c	0.26	-0.01	0.97	n.t.
Wong (2003)	0.23	n.t.	0.45	0.36
Average effect size	0.22	-0.13	0.55	-0.26
95% CI upper	0.47	0.07	0.97	0.19
95% CI lower	-0.03	-0.33	0.13	-0.70
<i>k</i>	17	6	11	7
<i>SD</i>	0.49	0.19	0.62	0.48

Note, n.t. = not tested in the particular study; n.a. = not applicable because the study did not provide sufficient data for the effect size calculations; *k* = number of samples contributing to the effect size calculation.

^a I could not calculate pre-to-post *d* values for Kubota's (2000) study because different measures were used for the pretest and posttest.

^b Shook (1999) contributed to the calculation of this value.

^c The *d* values were calculated from White's (1996) dissertation.