Language Learning 50:3, September 2000, pp. 417–528

Effectiveness of L2 Instruction: A Research Synthesis and Quantitative Meta-analysis

John M. Norris and Lourdes Ortega
University of Hawai'i at Manoa

This study employed (and reports in detail) systematic procedures for research synthesis and meta-analysis to summarize findings from experimental and quasi-experimental investigations into the effectiveness of L2 instruction published between 1980 and 1998. Comparisons of average effect sizes from 49 unique sample studies reporting sufficient data indicated that focused L2 instruction results in large target-oriented gains, that explicit types of instruction are more effective than implicit types, and that Focus on Form and Focus on Forms interventions result in equivalent and large effects. Further findings suggest that the effectiveness of L2 instruction is durable and that the type of outcome measures used in individual studies likely

John M. Norris and Lourdes Ortega, Department of Second Language Studies

We thank Craig Chaudron for prompting this research through discussions in his graduate seminar at the University of Hawai'i in the spring of 1997, and through a plenary he delivered at the seventh annual AAAL conference in Orlando, Florida, March, 1997 (we regret only that he failed to mention the enormity of such an undertaking). A second graduate seminar at the University of Hawai'i, taught by Cathy Doughty in the fall of 1997, provided the optimal context for early stages of the project, and initial findings were presented at the ninth annual AAAL conference, Stamford, Connecticut, March 7, 1999. We thank Cathy Doughty and Mike Long for supporting our work and, most importantly, for cultivating a domain of research worth synthesizing. The final version of this paper benefited from suggestions by the following individuals: three anonymous reviewers, the editor of *Language Learning*, Craig Chaudron, Cathy Doughty, and Mike Long.

Correspondence concerning this article should be addressed to John M. Norris at jnorris@hawaii.edu or Lourdes Ortega at lortega@hawaii.edu.

affects the magnitude of observed instructional effectiveness. Generalizability of findings is limited because the L2 type-of-instruction domain has yet to engage in rigorous empirical operationalization and replication of its central research constructs. Changes in research practices are recommended to enhance the future accumulation of knowledge about the effectiveness of L2 instruction.

Research on second-language acquisition over the past two decades has seen a proliferation of quasi-experimental and experimental studies that address the effectiveness of various instructional treatments in L2 classrooms (Doughty & Williams, 1998a) as well as in laboratory settings (Hulstijn, 1997). Indeed, a relatively well-defined research agenda appears to have emerged in L2 instruction research, since Long (1983) concluded that instruction makes a difference in L2 acquisition, when compared with naturalistic exposure. The principal focus of L2 instruction research has thus evolved from whether or not instruction makes a difference to what types of instruction are most effective for fostering second or foreign language learning in formal contexts (Doughty, 1991; Long, 1991a).

L2 type-of-instruction research to date has investigated the following general research questions:

1. Is an implicit or an explicit approach more effective for short-term L2 instruction? (e.g., Alanen, 1995; de Graaff, 1997; DeKeyser, 1995; Doughty, 1991; Ellis, 1993; Robinson, 1996b; Scott, 1989, 1990);

2. Can raising learners' metalinguistic awareness of specific L2 forms facilitate acquisition by fostering psycholinguistic processes of form-to-function mapping? (e.g., Fotos, 1993, 1994; Fotos & Ellis, 1991; Kubota, 1995b; Swain, 1998);

3. Is instruction that draws learners' attention to relevant forms in the context of meaning-focused lessons more effective than an exclusive focus on meaning and content? (e.g., Lightbown & Spada, 1990; functional-analytic instruction in Day &

Shapson, 1991; Harley, 1989; Lyster, 1994; and Leeman, Arteagoitia, Fridman, & Doughty, 1995; Williams & Evans, 1998);

4. Is negative feedback beneficial for L2 development, and if so, what types of feedback may be most effective? (e.g., descriptive research by Chaudron, 1977; Lyster, 1998; and experimental studies by Carroll, Roberge, & Swain, 1992; Carroll & Swain, 1993; Doughty & Varela, 1998; Kubota, 1994, 1995a, 1996; Long, Inagaki, & Ortega, 1998; Mackey & Philp, 1998; Nagata, 1993; Nobuyoshi & Ellis, 1993; White, 1991);

5. Is acquisition promoted more effectively when learners process the input in psycholinguistically relevant ways than when they experience traditional grammar explanation and practice? (e.g., Cadierno, 1995; VanPatten & Cadierno, 1993a, 1993b; VanPatten & Oikkenon, 1996; VanPatten & Sanz, 1995);

6. Is comprehension practice as effective as production practice for learning L2 structures? (e.g., DeKeyser, 1997; DeKeyser & Sokalski, 1996; Nagata, 1998; Salaberry, 1997).

These type-of-instruction studies share the theoretical premise that the goal of any instructional interventions should be to effect changes in learners' focal attention when they are processing the L2 (Sharwood Smith, 1993), so as to increase the likelihood that certain linguistic features are noticed (Schmidt, 1993, 1997) and eventually acquired, and to do this in efficient ways in terms of rate of acquisition and target-like levels of ultimate attainment. Hence, a central concern of such research is whether optimal L2 learning takes place through implicit or explicit cognitive processing of new material (N. Ellis, 1994; Leow, 1998b; Robinson, 1995b; Tomlin & Villa, 1994).

In the wider field of second language acquisition, it is a point of theoretical debate whether external efforts to "teach" L2 knowledge can truly impact on learners' developing L2 grammars.[1] Within this debate, L2 type-of-instruction research that subscribes

to the so-called weak interface position argues that certain instructional techniques, which contextualize the new L2 material within meaningful episodes in a manner that is relatively unobtrusive but salient enough for further cognitive processing, may help learners direct their attention to the relevant features in the input, and thus may expedite the acquisition process (see Doughty & Williams, 1998a; de Graaff, 1997; Sharwood Smith, 1981; Terrell, 1991). Such research is thus concerned with precisely how instructional interventions may best create a window of opportunity for external manipulation of learners' focal attention. Other approaches within L2 type-of-instruction research subscribe to the so-called strong interface position, investigating how declarative knowledge may be converted into implicit knowledge available for spontaneous L2 use (e.g., DeKeyser, 1997; McLaughlin, 1990; McLaughlin & Heredia, 1996).

Studies of instructional effectiveness have actualized the various concerns outlined above according to several descriptive models for types of L2 instruction. Long (1991b, 1997; Long & Robinson, 1998) has proposed that instructional options can be of three types, depending on whether instruction requires learners to focus on meaning, forms, or an integration of both meaning and forms. According to Long, instruction that is based on a focus on meaning posits that exposure to rich input and meaningful use of the L2 can lead to incidental acquisition of the L2 system. Instruction that expects learners to focus on forms in isolation (focus-on-forms or FonFS instruction) assumes that the target L2 forms can and need to be taught one by one in a sequence externally orchestrated according to linguistic complexity. Finally, instruction that seeks to make learners focus on forms integrated in meaning (focus-on-form or FonF instruction) capitalizes on brief, reactive interventions that, in the context of meaningful communication, draw learners' attention to formal properties of a linguistic feature which appears to cause trouble on that occasion, is learnable given the learner's internal developmental state, and is likely to be useful in future communication. Long (1991b, 1997; Long & Crookes, 1992, 1993) contends that FonF instruction is likely to be

more effective because it is consonant with what L2 researchers know about how second languages are acquired.

Spada (1997), on the other hand, proposes the term form-focused instruction (FFI) to characterize a wider range of instructional types that concur with theories of the role of consciousness and attention in L2 learning (Schmidt, 1993, 1997; Sharwood Smith, 1993), regardless of whether they are reactive or proactive or relatively obtrusive or unobtrusive. Such FFI interventions attempt to foster learners' shift of focal attention to particular forms within a meaningful context, but they may do so with a predetermined linguistic syllabus in mind, which is to be integrated into the otherwise content-based and meaning-oriented syllabus of the L2 classroom (see also arguments in Lightbown, 1998; Lyster, 1990).

Taking an intermediate position, Doughty and Williams (1998b) consider as definitional criteria for focus on form (FonF) instruction: (a) that learner engagement with meaning occur before attention to the linguistic code, possibly by ensuring that particular target forms are essential or at least natural for the completion of a task (as in Loschky & Bley-Vroman, 1993); (b) that an analysis of learner needs trigger the instructional treatment (following Long & Robinson, 1998), whether such analysis occurs reactively or proactively (departing here from Long & Robinson's definition and thus including some of the instructional types characterized as FFI in Spada, 1997); and (c) that learner focal attention is drawn to form briefly and overtly, that is, achieving a difficult balance between unobtrusiveness and salience (Doughty, 1997).

These models delineate the core dimensions of instructional treatments within L2 type-of-instruction research. Particular selections and combinations of related instructional features constitute more specific pedagogical techniques that have begun to be investigated in recent years. Well-known examples of these are implicit–inductive grammar teaching, traditional explicit grammar explanation, consciousness-raising activities, and dictogloss (all rule-based instructional types); recasts, enhanced output through provision of

clarification requests, garden path, models, and metalinguistic feedback (all feedback-based instructional types); flood, typographical input enhancement, and pre-emptive models (all input-based instructional types); and input-processing instruction and output practice (both practice-based instructional types).

As this research agenda has developed, it has also become more complex, with previously absolute questions about the effectiveness of various types of L2 instruction being redefined and stipulated according to possible moderator variables (see Doughty & Williams, 1998c; Ellis & Laporte, 1997; Hulstijn & de Graaff, 1994; Lightbown, 1998; Spada, 1997). L2 type-of-instruction research has become concerned with the relative effectiveness of particular instructional treatments when they are matched with specific learner characteristics, such as internal status of a learner's interlanguage, age, language aptitude, L1 background, etc. (see theoretical discussion in Chaudron, 1985; Lightbown, 1985, 1998; Pienemann, 1984, 1989; Robinson, 1995a; Skehan, 1998), and as they bring about the acquisition of specific L2 target features, for instance, simple versus complex forms (e.g., empirical work by Alanen, 1995; de Graaff, 1997; DeKeyser, 1995; N. Ellis, 1993; Robinson, 1996b).

Before the field can begin to systematically address the complex interactions of this developing research agenda, it is imperative to evaluate the findings that have emerged from L2 type-of-instruction studies to date. In the absence of consistent answers to fundamental questions about L2 instructional effectiveness, there is no cumulative context for situating new directions in research or for interpreting new findings. As Cooper (1998) has emphasized, "Given the cumulative nature of science, trustworthy accounts of past research are a necessary condition for orderly knowledge building" (p. 1).

Synthesizing Research on the Effectiveness of L2 Instruction

Given broad categorical similarities among experimental and quasi-experimental studies of instructional effectiveness, L2

type-of-instruction research appears to have evolved into a research domain addressing a homogeneous, if general, set of research problems, summarized here in the form of two overarching questions:

1. How effective is L2 instruction (versus simple exposure or meaning-driven communication)?

2. What is the relative effectiveness of different types of L2 instruction?

Primary research (i.e., investigations that gather data and conduct analyses on individual study samples) within the domain has produced increasing amounts of data in response to these and ancillary questions. However, regardless of how big the sample size or complex the design, no single investigation of the effectiveness of L2 instruction can begin to provide trustworthy answers (i.e., indications of consistent patterns to be observed within particular variables and upon which interpretations may be based with high degrees of probability). Individual study findings are too easily attributable to chance variability as well as to idiosyncrasies in design, analysis, sampling error, research setting, etc. (Cooper, 1998; Light & Pillemer, 1984; Taveggia, 1974). Instead, to search for answers within the domain, mounting findings from primary research are best utilized as evidence in secondary research, which takes stock of a given domain in two ways: (a) by gathering and weighing the available evidence offered by results from all primary studies addressing a common research problem; and (b) by assessing the consistency or statistical trustworthiness of answers offered by the preponderance of evidence gathered from multiple research contexts. Thorough secondary research should therefore serve as a kind of watershed point in cumulative scientific endeavor, summarizing what has come before and indicating what remains to be done (Rosenthal, 1991).

Reviews of research on L2 instruction. Substantial cumulative secondary research has been undertaken within the L2 type-of-instruction literature, appearing in review articles, chapters, or

books that summarize the state of research efforts and findings (Chaudron, 1988, 1998; DeKeyser, 1994; Doughty, 1998; Doughty & Williams, 1998a; N. Ellis, 1994; N. Ellis & Laporte, 1997; R. Ellis, 1994, 1998; Harley, 1988, 1994; Hulstijn, 1997; Kasper, 1998; Krashen, 1999; Lightbown, 1985, 1998; Long, 1983, 1988, 1991b; Long & Robinson, 1998; Schmidt, 1993; Spada, 1997; Truscott, 1996; VanPatten 1988, 1994; Williams, 1995). Such cumulative secondary work has generally adopted either a narrative or a vote-counting approach to research review (Light & Pillemer, 1984), each of which has serious limitations as a means for accumulating and synthesizing scientific knowledge.

For several reasons, narrative reviews may not provide the most accurate picture of the state of cumulative knowledge available from primary research findings. One weakness with such reviews is incomplete identification and recovery of relevant primary research. Thus, separate reviews of the same question may draw conflicting conclusions about the state of findings (see, e.g., conclusions drawn by Krashen, 1999, versus those in Spada, 1997), owing to inconsistent sampling of primary studies. In addition, as Light and Pillemer (1984) have noted, "the personal beliefs of a reviewer can play a role in resolving disparate findings" (p. 5). Thus, two researchers may interpret the same study findings in very different ways because they are using different evaluative criteria. Finally, the most prevalent problem in narrative reviews arises when reviewers base their conclusions on the conclusions drawn by primary researchers, which, as Long (1983) among others has demonstrated, may have little to do with what the research data actually showed (see also Dubin & Taveggia, 1968; Rosenthal, 1991).

To ameliorate the ad hoc nature of narrative reviews, vote-counting reviews (Light & Smith, 1971; Bushman, 1994) begin by identifying all of the studies within a domain that have addressed a particular research question. Evidence provided by each of these primary studies in the form of statistically significant or nonsignificant findings is then utilized to "cast a vote," either supporting (statistically significant in the hypothesized direction), not

supporting (not statistically significant), or contradicting (statistically significant in the opposite direction) a particular hypothesized answer to the question. Based on a tally of the votes, conclusions are drawn about what the evidence seems to suggest. Although derived directly from research data, and therefore not dependent on the conclusions drawn by primary researchers, the vote-counting review may nevertheless result in incorrect interpretations of what the evidence actually has to say. Namely, the simple conclusion of "statistically significant" or "not statistically significant" may all too often mask an actual observed effect or relationship, because statistical significance is directly dependent on the sample sizes within primary studies. Indeed, two studies observing exactly the same effect may come to contradictory conclusions merely because of differences in sample sizes (see related discussions in Bangert-Drowns, 1986; Carver, 1993; Cohen, 1992; Harlow, Muliak, & Steiger, 1997; Meehl, 1991; Rosenthal, 1994).[2]

Both narrative and vote-counting reviews are also limited in that they provide no information about the magnitude of an effect, the strength of a relationship, or the importance of a finding observed within a group of studies. Light and Pillemer (1984) note that "even if every one of 30 studies in a review reports findings that are statistically significant, a vote count does not tell us whether they are large enough to matter in practice" (p. 75). Finally, neither procedure relates anything about the statistical trustworthiness (e.g., in terms of the standard error associated with observations) of an overall finding (Cooper, 1998; see also methods for improving vote-counting procedures in Bushman, 1994; Hedges & Olkin, 1980).

Research synthesis and meta-analysis. Research synthesis has developed recently into a science in its own right, providing secondary researchers with replicable methods that produce verifiable findings (Cooper, 1998; Cooper & Hedges, 1994a; Light & Pillemer, 1984). Such methods enable researchers to summarize the state of cumulative knowledge within a domain by adhering to organizational principles for the sampling of primary research

studies, the evaluation and classification of substantive and methodological study features, and the analysis and interpretation of study findings (Cooper & Hedges, 1994b). Depending on a domain's maturity, research synthesis may be more exploratory and descriptive, reviewing scientific processes within the domain (e.g., Glass, 1976), or more confirmatory, inferring causal relationships from the body of research data (e.g., Cook, 1992).

To enable precise analysis and interpretation of primary research findings, particular focus within the research synthesis literature has been given to methods for quantitative meta-analysis (Cooper & Hedges, 1994a; Hedges & Olkin, 1985; Hunter & Schmidt, 1990; Rosenthal, 1991; Wolf, 1986). The fundamental premise of meta-analytic interpretation is that all available primary research findings, typically in the form of basic descriptive statistics, provide data for estimating the overall observed finding about a given treatment or condition across studies sharing a research focus. Meta-analysis enables this estimate by treating primary research data according to a common scale. Thus, findings from individual studies are converted to comparable values by estimating the magnitude of an observed relationship or effect, typically referred to as the *effect size*. Several types of effect size estimates have been developed (Kirk, 1996; Rosenthal, 1994; Wolf, 1986), although the most commonly reported is the standardized mean difference, which can be directly interpreted as the difference between two research groups (i.e., treatment and control groups) in standard deviation units.

Once the data from a range of related studies have been standardized, study findings can be combined to produce an average effect size (e.g., summarizing a treatment's effectiveness across studies), or they can be compared to investigate differences in study findings and relate these differences to study features (e.g., different dependent variables, moderator variables, etc.). Combining and comparing effect sizes gives a much more accurate picture of the actual effects observed within primary research than does a vote-count of statistically significant findings. Interpretation of effect sizes is not endangered by differences in sample

sizes among studies (although effect sizes may be weighted according to differences in study sample sizes; see Shadish & Haddock, 1994), and effect sizes can be interpreted without the use of statistical significance tests (Cohen, 1990, 1997). Meta-analysis also incorporates statistical procedures for estimating the extent to which interpretations are trustworthy, by considering the frequency and consistency of observed effects for a given variable of interest (i.e., in the form of standard errors and confidence intervals) across a domain of studies (see Rosenthal, 1995; Shadish & Haddock, 1994; for more on meta-analytic methods, see Bangert-Drowns, 1986; Cooper & Hedges, 1994a; Rosenthal & Rubin, 1986).

The current study. We undertook a synthesis of primary research on the effectiveness of L2 instruction, incorporating systematic procedures to survey the research domain and quantitative meta-analytic techniques to summarize and interpret study findings. Although not the first study within applied linguistics to utilize meta-analytic techniques (see Ross, 1998; Sahari, 1997), this is the first study to synthesize L2 instructional research using meta-analysis. The research domain was defined as all published experimental or quasi-experimental primary research investigating the effectiveness of L2 instructional treatments. The domain was limited to experimental and quasi-experimental studies, owing to the focal research questions (asking about the effect of a treatment) and the likelihood of retrieving quantitative data within such research.

The kinds of questions within the domain that had accumulated substantial primary research data motivated the research questions for the current study. Ellis and Laporte (1997) have correctly observed the following about this domain:

> Formal instruction is too catch-all a category, as method is too poorly defined a term (Long, 1991[b]) to allow much sense from putting all of these studies in the same meta-analysis and reviewing them together. We are only just beginning to gather a sufficient quantity of studies to allow us finer categories of comparison so that we can investigate the effects of particular methods of instruction with particular

> content and focus on particular outcome measures (fluency vs. accuracy, comprehension vs. production, etc.) in particular learners of particular learning styles at particular stages of development (e.g., Long, 1988). (p. 66)

Although the research domain has only produced small numbers of studies addressing such specific interactions among variables, it has accumulated many studies investigating more general questions of overarching interest to the domain. In other fields, meta-analyses have been conducted on research questions as general as the overall effects of psychotherapy (Smith & Glass, 1977; Rosenthal, 1983), the relationship between gender and cognition (Rosenthal & Rubin, 1982), the effect of rewards on intrinsic motivation (Cameron & Pierce, 1994), and the effect of homework on academic achievement (Cooper, 1989). In the end, the use of quantitative meta-analytic procedures in conjunction with a systematic research synthesis should produce a precise, replicable, and verifiable account of whatever the state of cumulative knowledge may be about a research question within a given domain.

Common to all L2 type-of-instruction studies is the investigation of different treatments that may be categorized according to the manner in which instructional delivery focuses learner attention on target L2 features. For present purposes, analysis of the effectiveness of different types of L2 instruction proceeded according to the model of instructional delivery categories proposed by Doughty and Williams (1998b, 1998c).

Accordingly, the two general research questions (RQs) identified above provided the primary focus for the current research synthesis and meta-analysis:

1. How effective is L2 instruction overall and relative to simple exposure or meaning-driven communication?

2. What is the relative effectiveness of different types and categories of L2 instruction?

In addition, three further questions, closely related to the interpretation of instructional effectiveness, were also addressed:

3. Does type of outcome measure influence observed instructional effectiveness?

4. Does length of instruction influence observed instructional effectiveness?

5. Does instructional effect last beyond immediate postexperimental observations?

Questions 1 through 5 were addressed by surveying the literature for relevant primary research, categorizing studies according to a model of L2 instruction, and quantitatively summarizing observed study findings. Average effect sizes were calculated to measure the magnitude of instructional effectiveness, and confidence intervals were estimated to gauge the statistical trustworthiness of observed effects (Rosenthal, 1991). This portion of the current study aimed to clarify exactly what research to date has shown about the effectiveness of general categories of L2 instruction. We hoped that pursuing this objective would reveal the present state of knowledge about fundamental questions and illuminate variables in need of systematic investigation in the future, and would also establish baseline data for the interpretation of future primary research findings. The current study also addressed a final question:

6. To what extent has primary research provided answers to these questions?

Question 6 was addressed by surveying the body of primary studies and summarizing the state of research design, analysis, and reporting. It was the objective of this portion of the study to assay the manner in which L2 type-of-instruction investigations have been conducted, to provide insight into areas in need of empirical attention, and to encourage improved research practice.

Method

The Literature Search

After specifying the research domain and formulating the research questions to be addressed in the current synthesis, the body of relevant study reports was identified through a principled, replicable, and exhaustive search of literature. In the same way that the individual study participant supplies data for primary research, the individual study supplies data for research synthesis. However, unlike primary research, where individuals are representatively or randomly sampled from a population to which findings will be generalized, secondary research attempts to sample the entire population of primary research studies in order to summarize the state of existing findings within a domain (Cooper, 1998).

To access the initial body of literature, key- and subject-word searches were conducted within the Educational Resources Information Center (ERIC) computer database; searches utilized the following words and word combinations: (a) focus on form, (b) form-focus(s)ed, (c) effect(s) (and effectiveness) of second (and foreign) language instruction, (d) negative feedback and second (and foreign) language instruction, (e) grammar instruction (and teaching), (f) explicit learning (and instruction), (g) implicit learning (and instruction), (h) consciousness raising and language learning (and instruction), (i) error correction and language learning (and instruction), (j) input processing, (k) recasts, (l) models, and (m) instructed second language acquisition (and learning).

Subsequently, several other search techniques were utilized. Back issues of 14 academic journals were browsed for relevant study reports.[3] Reference sections from a number of reviews of the research domain were consulted (Chaudron, 1988, 1998; DeKeyser, 1994; Doughty, 1998; Doughty & Williams, 1998a; N. Ellis, 1994; N. Ellis & Laporte, 1997; R. Ellis, 1994, 1998; Harley, 1988, 1994; Hulstijn, 1997; Kasper, 1998; Krashen, 1999; Lightbown, 1985, 1998; Long, 1983, 1988, 1991b; Long & Robinson, 1998;

Schmidt, 1993; Spada, 1997; Truscott, 1996; VanPatten 1988, 1994; Williams, 1995). Finally, reference sections from each retrieved study report were cross-checked for additional study reports.[4] After identifying this large initial "net" of potentially relevant study reports, all studies were retrieved through library services, ERIC reproduction services, direct purchase from publishers, or personal requests from individuals with access to the particular sources.[5]

It is important to point out that no attempt was made in the current study to retrieve the so-called "fugitive" literature (e.g., unpublished papers, dissertations and theses, conference presentations). Rosenthal (1994) maintains that the most comprehensive synthesis of the state of knowledge about a research question should include not only published sources but also hard to find "fugitive" sources. There are generally two reasons for including the fugitive literature. First, it may be that some research reports that simply have not reached a published forum, for any number of reasons, could nevertheless provide further primary data to increase the accuracy of a synthesis of the overall findings within a research domain.

The second and more important reason for retrieving fugitive sources is to avoid the very real risk of incurring the "file-drawer" problem in research synthesis (Rosenthal, 1979a). The file-drawer problem issues from the well-attested fact that studies reporting statistically significant findings tend to be accepted for publication over studies reporting no statistically significant findings (e.g., Atkinson, Furlong, & Wampold, 1982; Begg, 1994; Cooper, DeNeve, & Charlton, 1997; Greenwald, 1975; Lipsey & Wilson, 1993). Unfortunately, as Cooper (1998) has pointed out, "This bias is present in the decisions made by both reviewers and primary researchers" (p. 54). As a direct result of such publication bias, it is assumed that a large number of studies exist in the file drawers of researchers who, having "failed" to reach statistical significance with a particular study, have filed the results away and tried again with a new study. Naturally, this practice on the part of researchers, journal editors, and reviewers can propagate a very

biased view of the actual state of cumulative scientific knowledge in a given research domain.

For the purposes of the current study, however, the fugitive literature was not consulted, as the primary goals of this research synthesis were to investigate the study characteristics as well as the study findings of the body of accessible, and therefore most influential, research literature. Readers should be fully aware, then, that there is most likely a serious publication bias influencing the results of the quantitative meta-analysis of study findings reported here. However, this approach enabled an accurate synthesis of exactly those findings from those studies that are published and reported, and that therefore in many ways define this research domain. Additionally, graphic techniques were utilized in order to assess the probable extent of publication bias within the body of study reports (see Results section).

Criteria for Inclusion in the Research Synthesis

Over 250 potentially relevant study reports were retrieved from the initial literature search. Both researchers reviewed each report to determine the actual relevance of the study to the research domain and current research questions. To be included in the synthesis, a study report had to meet all of the following criteria:

1. The study was published between 1980 and 1998. Although several studies published prior to 1980 were retrieved in the literature search, they were judged inappropriate because of reporting infelicities, uncharacteristic research designs and analyses, and a dissimilar focus in research questions.[6] Studies published after 1998 were not available in their published form at the time of the current synthesis.

2. The study had a quasi-experimental or experimental design. Only those studies experimentally investigating the effectiveness of particular L2 instructional treatments could contribute data for the calculation of effect sizes.

3. The independent variable(s): (a) constituted an adequately defined and reported instructional treatment (which could be treated as an independent variable and compared with other studies), and (b) targeted specific forms and functions, either morphological, syntactic, or pragmatic (as the theoretical underpinnings of type-of-instruction research focus on the acquisition of rule-governed aspects of the L2, with special attention to morphology, syntax, and, more recently, pragmatics).[7]

4. The dependent variable(s) was a measure of language behavior related to the specific structures targeted by independent variables. Only studies reporting such outcome measures could provide interpretable findings about the effectiveness of particular instructional treatments.

In general, then, a number of studies identified within the initial literature search were not included in the current research synthesis for the following reasons:

1. Studies utilized descriptive or correlational designs (e.g., Lightbown & Spada, 1990).

2. Instructional treatment did not focus on the learning of a specific form(s) (e.g., Robb, Ross, & Shortreed, 1986).

3. The target of instruction was phonology or lexis (e.g., Tomasello & Herron, 1989).

4. Dependent variables did not measure the impact of instructional treatments on the learning of specific structures (e.g., Gass & Varonis, 1994).

We do not wish to suggest that these studies or the associated RQs and designs are inadequate or inappropriate. Rather, they were simply not related to the focal RQs for the current synthesis. After evaluating all of the retrieved literature according to the inclusion criteria above, 77 individual study reports were retained as the body of research for this synthesis (these study reports are identified with single and double asterisks in the reference section).

It should be noted that the literature on quantitative meta-analysis also typically recommends establishing research quality criteria for inclusion decisions (e.g., Petrosino, 1995; Rosenthal, 1995; Wortman, 1994). Thus, for example, studies may be excluded on the basis of a number of threats to internal and external validity (e.g., no researcher blind, no randomized sampling of study participants, no control groups; see 33 specific threats in Cook and Campbell, 1979). However, an inclusive approach was adopted in the current synthesis, and no such decisions were made based on the validity of the primary research reported. Methodological and substantive inconsistencies within the research domain tend to be so prevalent that an inclusive approach was necessary to make this initial synthesis of the domain at all possible. In addition, one focus of the synthesis was to summarize and evaluate the range of research practices applied within the domain, making an inclusive approach all the more necessary.

Coding Study Reports

After identifying the body of research literature meeting inclusion criteria, we coded and categorized the resulting 77 study reports according to a variety of study features. Coding was undertaken in order to: (a) describe systematically how researchers have investigated the effectiveness of L2 instructional techniques; (b) clarify the variables of interest to the research domain by using categories generic to all studies, regardless of definitional discrepancies among primary researchers; (c) classify studies according to similarities among research variables (e.g., instructional treatments, outcomes measures); and (d) identify research findings in the form of data appropriate for inclusion in the quantitative meta-analysis. To diminish the potential effects of expectancy bias, which can lead to arbitrary interpretations of findings by research synthesists (Cooper, 1998), the coding of study reports ensued after definitions for study features had been established. Additionally, coding proceeded according to a series of stages, and checks on researchers' judgments were included at

each of these stages to ensure the reliability of the process (Yeaton & Wortman, 1993). The development of coding categories and the reliability of codings are explained in detail below.

Following the research synthesis literature (e.g., Glass, McGaw, & Smith, 1981; Stock, 1994), definitions for substantive and methodological categories of study features were established in the following manner. The two researchers independently coded four representative study reports for salient substantive and methodological features, then discussed these features and agreed upon corresponding definitions. The features and definitions were further reviewed by colleagues and research domain experts, and their feedback was incorporated.[8] Based on these initial codings and discussions, a coding book (Stock, 1994) was created to ensure the systematic review of all study reports according to salient study features (see definitions below).

Using the coding book, both researchers independently coded a subsample of 20% (n = 16) of the 77 study reports for all study features. The simple agreement ratio (Orwin, 1994) between the two researchers was 0.88 for this initial round of coding. Disagreements were discussed and resolved, and study feature definitions were refined. Coding of low-inference features (e.g., duration of treatment, timing of tests, sizes of samples, etc.) for the remainder of the 77 study reports was subdivided between the two researchers, with the first researcher coding research findings and reporting characteristics, the second researcher coding instructional treatment subtypes, and both researchers coding methodological design features.

Following these low-inference codings, a last round of high-inference decision making was undertaken. In this final round, each researcher independently coded all 77 study reports for the following categorical decisions: (a) Were instructional treatments best characterized as focus-on-form or focus-on-forms? (b) Were instructional treatments explicit or implicit? (c) Were outcome measures based on metalinguistic judgments, selected responses, constrained constructed responses, or free constructed responses? (d) Which findings were available for quantitative meta-analysis?

An overall agreement ratio of 0.95 was observed for these decisions, and disagreements were resolved through review and discussion of the original study reports.

In summary, the 77 study reports were coded for various features in order to provide a basis for describing the research domain as well as to identify categories and data for further synthesis. The final coding categories are described and defined in the next section, to provide the reader with a basis for understanding discussions in subsequent sections. Additionally, for substantive study features, more detailed inter-coder agreement analysis is provided in light of the high-inference nature of the coding categories (Yeaton & Wortman, 1993).

Substantive features. Substantive features consisted of those independent and dependent variables through which primary researchers operationalized their investigations of the effectiveness of L2 instruction. Such variables thus comprised the primary focus of the research domain, providing interpretive links between hypothesized relationships and empirical observations. For several reasons, coding of these substantive features was a high-inference procedure. First, reporting of the exact operationalization of instructional treatments (the independent variables) as well as outcome measures (the dependent variables) is extremely variable across study reports within L2 type-of-instruction research. Second, not only do primary researchers disagree on the exact attributes of various categories of L2 instructional treatments, but, as Doughty and Williams (1998a, p. 3) have pointed out, they also utilize differing terminology for describing instructional treatments and outcome measures. We therefore took care to code particular types of independent and dependent variables according to the following generic categorical definitions and on the basis of the evidence supplied in study reports.

Independent variables. Particular instructional treatments were first classified according to whether they could be considered explicit or implicit and whether they attempted to shift learners' attentional focus onto form (FonF), forms (FonFS), or meaning (FonM) (Long, 1991b; Long & Robinson, 1998). It should be reiterated

here that a compromise definition for FonF versus FonFS treatments was adopted, following Doughty and Williams (1998b, 1998c) rather than the more restrictive definition of FonF in Long and Robinson (1998) or the more inclusive definition of FFI in Spada (1997).

Following DeKeyser (1995), an L2 instructional treatment was considered to be *explicit* if rule explanation comprised part of the instruction (in this first sense, explicit designates deductive and metalinguistic) or if learners were directly asked to attend to particular forms and to try to arrive at metalinguistic generalizations on their own (in this second sense, explicit designates explicit induction).[9] Conversely, when neither rule presentation nor directions to attend to particular forms were part of a treatment, that treatment was considered *implicit*. Thus, for example, Scott's (1989) flood treatment was classified as implicit, given the fact that the teacher read aloud to students short episodes of a story over six consecutive class periods with the only requirement being that they listen to the stories and answer comprehension questions. Scott's (1990) flood treatment, on the other hand, was classified as explicit because the treatment explicitly induced students to pay attention to the target forms by telling them that the stories would contain many relative clauses and by asking them to listen for these forms. Similarly, both the garden path and the pre-emptive modeling treatments in all garden path studies (Ellis, Rosszell, & Takashima, 1994; Kubota, 1995a; Tomasello & Herron, 1988) were classified as explicit because grammar rules were explained at some point in both instructional types. By contrast, both the recast and the pre-emptive modeling treatments in Long, Inagaki, and Ortega (1998) were classified as implicit because neither grammar explanations nor instructions to attend to L2 forms were given in these treatments.

The characterization of an instructional treatment as FonF, FonFS, or FonM instruction was a higher inference decision than was the implicit/explicit distinction. Adhering to the definition of focus-on-form in Doughty and Williams (1998b, 1998c), the following solution was agreed upon for the current synthesis. An

instructional treatment was classified as FonF if there was evidence that an integration of form and meaning was addressed via any of the following strategies: (a) designing tasks to promote learner engagement with meaning prior to form; (b) seeking to attain and document task essentialness or naturalness of the L2 forms; (c) attempting to ensure that instruction was unobtrusive; (d) documenting learner mental processes ("noticing"). In addition, many FonF studies also presented evidence of: (e) selecting target form(s) by analysis of learners' needs; or (f) considering interlanguage constraints when choosing the targets of instruction and when interpreting the outcomes of instruction. In coding study reports, each of the researchers independently looked for evidence of any of these strategies in the description of an instructional treatment and then decided whether the treatment qualified as FonF. Instructional treatments were classified as FonFS when the following two conditions were in evidence: (a) none of the four strategies (a)–(d) above could be identified; and (b) learner attention was nevertheless focused in some way on the particular structure targeted for learning. Thus, for example, input-processing treatments (Cadierno, 1995; VanPatten & Cadierno, 1993a, 1993b; VanPatten & Oikkenon, 1996; VanPatten & Sanz, 1995) were classified as FonF because the comprehension practice activities delivered in these treatments were reportedly designed to integrate learner engagement in meaning with a focus on formal aspects of the L2 targets (Cadierno, 1992). The traditional practice treatments in these studies, on the other hand, were classified as FonFS because learners engaged in production activities ranging from mechanical to more communicative drills, and an integration of form and meaning was not built into these activities. In contrast with the input processing studies, both the comprehension and production treatments in DeKeyser and Sokalski (1996), Nagata (1998), and Salaberry (1997) were classified as FonFS, given that none of these studies presented evidence that an integration of form and meaning was sought, none discussed any of the four FonF strategies listed above, and both types of treatment (as well as

subsequent outcome measures) involved the manipulation of minimally contextualized forms.[10]

Finally, any experimental treatment or condition which involved exposure to the L2 targets or experience with the L2 tasks, but which did not involve an attempt at effecting shifts in learner attention to L2 target structures, was coded as FonM. Table 1 shows the frequency of decisions, intercoder agreement ratios, and kappa coefficients for coding of independent variables.[11]

In addition to coding independent variables according to the explicit/implicit and FonF/FonFS/FonM distinctions, and to characterize particular approaches to instructional treatments that have received attention in research on the effectiveness of L2 instructional techniques, all independent variables were further classified into subtypes. This further classification involved the relatively low-inference identification of some instructional treatment subtypes well known in the domain: (a) flood; (b) enhancement; (c) recasts; (d) consciousness raising; (e) input processing; and (f) garden path. In addition, the following further subtypes were identified with descriptive labels: (g) traditional explicit; (h) traditional implicit; (i) input practice; (j) output practice; (k) metalinguistic feedback; (l) metalinguistic task-essentialness; (m) rule-oriented forms-focused; (n) rule-oriented FonF; (o) other implicit; (p) pre-emptive modeling; (q) compound FonF; (r) corrective models; and (s) form-experimental. These descriptive labels were created in an effort to characterize consistently across studies those specific techniques that shared common instructional features. Appendix A shows the substantive features for type and subtype of instruction coded for each experimental condition across the 49 studies included in the quantitative meta-analysis.

Dependent variables. Given the fact that effectiveness of instructional treatments was assessed and interpreted by primary researchers according to changes in what study participants were able to do or to demonstrate in the L2 on various outcome measures, we decided to code these dependent variables according to the type of activity or response required of the learner (i.e., the characteristic most directly linked to eventual interpretations). After

Table 1

Coding Reliability for Substantive Study Features

Study features (n = 9)	Coding frequency	Agreement ratio	Cohen's kappa
Independent variables			
(n = 5)	143	0.93	0.91
Dependent variables			
(n = 4)	141	0.98	0.97

Note: Frequency of coded variables exceeds the number of study reports owing to the presence of multiple independent and dependent variables within individual primary research studies.

initial review of the research domain, four general response types were identified. Measures were coded as *metalinguistic judgments* if the research participant was required to evaluate the appropriacy or grammaticality of L2 target structures as used in item prompts (e.g., grammaticality judgment tasks). *Selected response* measures required participants to choose the correct response from a range of alternatives, typically either in answer to comprehension questions based on the use of the target L2 form(s) or in order to complete a sample segment of the target language with the appropriate target form(s) (e.g., multiple choice tests providing four options in verbal morphology). Outcome measures were coded as *constrained constructed response* if they required the participant to produce the target form(s) under highly regulated circumstances, where the use of the appropriate form was essential for grammatical accuracy to occur. Constrained constructed response measures required learners to produce L2 segments ranging in length from a single word up to a full sentence, but all such measures were designed with the intent to test L2 ability to use the particular form within a highly controlled linguistic context (e.g., sentence combining with relative pronouns). *Free constructed response* measures were those measures that required participants to produce language with relatively few constraints and with meaningful communication as the goal for L2 production (e.g., oral interviews, written compositions). Use of the target

form(s) in free constructed response measures was typically not induced or required for communication, and instances of the form were tallied by primary researchers in order to interpret the effectiveness of instructional treatments. An overall agreement ratio of 0.98 was observed between the two researchers in the current study for coding of dependent variables according to these features (see Table 1). Appendix A shows the dependent variable types associated with each of the 49 studies included in the quantitative meta-analysis.

Methodological features. All study reports were also coded for a range of methodological features in order to demonstrate the extent to which such information is adequately reported in primary research, and to provide an overview of methodologies from which to determine characteristics desirable in future research (Lipsey, 1994). Methodological features coded in the current synthesis included: (a) learner populations; (b) instructional settings; (c) research designs; and (d) statistical analyses. The particular features are presented along with summary data in the Results section (see Tables 2 through 6).

Where sufficient information was provided by primary researchers, the coding of methodological features was very low inference, typically involving a series of dichotomous decisions. However, it quickly became apparent that reporting was incomplete in a number of the studies. For the purposes of characterizing the ways in which primary researchers in the domain go about reporting their study methodologies, incomplete reporting posed no problems. However, for synthesizing an accurate picture of how researchers go about investigating the effectiveness of L2 instruction, incomplete reporting introduced a potential bias. Furthermore, for the purposes of summarizing research findings across study reports in quantitative meta-analysis, incomplete reporting of outcomes excluded particular studies. Different ways of dealing with the problem of incomplete reporting have been suggested (Cooper, 1998; Pigott, 1994), including the substitution of average or predicted values, and follow-up contact with primary researchers. However, in the current synthesis, given the exploratory

nature of the meta-analysis and the emphasis on describing the state of available research findings, it was decided that missing values should simply be incorporated as such.

The Quantitative Meta-analysis

To demonstrate exactly what available research data have to say about how effective different L2 instructional treatments may be, a quantitative meta-analysis of observed study findings was undertaken. In selecting from the different effect size estimates, Rosenthal (1994) recommends employing "*d*-type effect size estimates when the original studies have compared two groups so that the difference between their means and their within-group *S*'s or σ's are available" (p. 236). Given the designs adopted by most primary researchers within L2 type-of-instruction studies, Cohen's (1977) *d*-index was selected as the most appropriate effect size estimate. Calculating Cohen's *d* produces a standardized mean difference for any contrasts made between two groups within a primary research study. This effect size can be interpreted as the magnitude of an observed difference between two groups in standard deviation units. Unfortunately, although the APA Guidelines (American Psychological Association, 1994, p. 18) encourage primary researchers to report effect sizes, this was rarely the case within the current domain. Thus, secondary analysis of primary research data was called for in all studies but one (Master, 1994) to derive Cohen's *d* values.

Estimating d. Calculating Cohen's *d* requires only the most fundamental descriptive statistics: group sample sizes, and dependent variable means and standard deviations of the two groups being contrasted (typically, scores of a treatment group and a control group on some outcome measure). Equation 1 shows this effect size formula (adapted from Rosenthal, 1994, p. 237):

$$d = \frac{mean_e - mean_c}{S_w} \quad (1)$$

where $mean_e$ is the mean of the experimental or treatment group, $mean_c$ is the mean of the control or comparison group, and S_w is the pooled standard deviation of the experimental and control groups (see discussion below on defining contrasts). The pooled-between-groups standard deviation is calculated as follows (adapted from Hunter & Schmidt, 1990, p. 271):

$$S_w = \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2}{(n_1 - 1) + (n_2 - 1)} \tag{1.1}$$

where n is the sample size of either group and S is the standard deviation of either group. Although the standard deviation of the control group may be used in the denominator of Equation 1 (Cohen, 1977), the pooled-between-groups standard deviation was selected for the current study because the standard deviation of any single group was considered particularly susceptible to sampling error, due to the small sample sizes of most studies within the research domain. Cohen's d was calculated using Equations 1 and 1.1 for any study within the current synthesis in which the sample sizes, dependent variable means, and dependent variable standard deviations were reported for two study groups being compared.

Unfortunately, it was frequently the case that primary researchers did not report the basic descriptive statistics necessary for this calculation of d. However, d may also be derived from other information occasionally included in study reports, utilizing the values reported when a test of significance has been derived through inferential statistics. For the current research synthesis, the following two formulas were used (adapted from Rosenthal, 1994, p. 238):

$$d = \frac{t(n_1 + n_2)}{\left(\sqrt{df_{error}}\right)\left(\sqrt{n_1 n_2}\right)} \tag{2}$$

where t is the value reported in a t test comparison, n is the sample size of either group, and df_{error} is the degrees of freedom error term, and

$$d = \frac{\left(\sqrt{F}\right)\left(n_1 + n_2\right)}{\left(\sqrt{df_{error}}\right)\left(\sqrt{n_1 n_2}\right)} \quad (3)$$

where F is the value reported in an analysis of variance (ANOVA) (or related statistic). For any studies reporting exact values for t or F comparisons between two groups, as well as group sample sizes, d was calculated using Equation 2 or Equation 3. It should be noted here that, for any single contrast between two groups within a single study, calculating Equations 1, 2, and 3 would result in an identical effect size estimate d.[12]

In summary, the 77 study reports were surveyed by both researchers for sufficient data to enable calculation of Cohen's d using any of the three equations above. Thus, effect sizes were calculated and included in the quantitative meta-analysis for any study reporting sample sizes and: (a) means and standard deviations; or (b) individual scores on outcomes measures for all study participants; or (c) between-groups exact t or F values (with df numerator equal to 1). Of the 77 study reports, 45 (58%) were found to report sufficient interpretable data for calculating Cohen's d (these studies are indicated by two asterisks in the reference section).[13]

For purposes of clarity, it should be pointed out that the number of published study reports (n = 45) contributing data to the quantitative meta-analysis does not correspond to the number of independent experiments identified. There are several reasons for this. In some cases, more than one study report was based on a single experimental study. In other cases, multiple experimental studies using unique population samples were presented in a single study report (both cases are noted in the References section). To distinguish clearly among study reports and investigations, for any single experimental study that appeared in multiple study reports, one of these reports was selected to represent the data consistently, so that the same data did not appear more than once within the quantitative meta-analysis. For data from multiple experiments presented within a single study report, each unique

sample experiment was labeled independently (i.e., study 1, study 2), and these labels were maintained throughout the meta-analysis to consistently identify the data. Overall, then, 77 study report publications were retrieved that met the criteria for inclusion in the research synthesis. Within these 77 study reports, 78 unique sample studies were identified. Of these unique sample studies, 49 contributed sufficient data for inclusion in the quantitative meta-analysis, and these 49 unique sample studies were published in 45 study reports.

Defining contrasts for effect size calculation. From the meta-analytic perspective, the ideal primary research design is one in which a single experimental condition is contrasted with a single control condition on a single dependent variable. Data from such a design form the basis for a perfect analysis of the effectiveness of a treatment versus no treatment, in the form of a single effect size estimate that can concisely represent the finding of the study. Indeed, there is much to be said in favor of such simplicity in research design (and this point is therefore revisited in the Discussion section).

In the current meta-analysis, however, defining appropriate contrasts was less straightforward, owing to inconsistencies in reporting and to characteristics of research designs within the domain. To determine exactly which contrasts between which research groups would provide values for calculating study effect sizes, several issues had to be dealt with for a number of study reports, including: (a) Research designs did not include true control groups; (b) Studies did not consistently report pre-experimental values on all dependent variables; (c) Outcomes in individual studies were measured on more than one dependent variable; (d) More than one independent variable (instructional treatment) was investigated in the same study.

To include in the meta-analysis quantitative findings representing the widest range of instructional treatment types reported in the research domain, particular values to be contrasted in effect size calculations were determined according to the following decisions:

1. For studies reporting data on one or more treatment groups and a true control group (i.e., a group receiving neither instruction nor exposure related to the target structure except in pre- and post-tests) (n = 10), d was calculated by contrasting each experimental group with the control group on the immediate post-test.

2. For studies reporting data on one or more treatment groups and an instructional comparison group (i.e., a group receiving nonfocused exposure to the structures being taught in the experimental condition) (n = 20), d was calculated by contrasting each experimental group with the comparison group on the immediate post-test.

3. For studies that assumed zero prior L2 knowledge among learners (e.g., in artificial language studies) and that therefore did not involve a control group (n = 4), the instructional condition with the least attention-focused treatment was selected as the baseline comparison group (i.e., treatments involving the processing of experimental input under largely incidental conditions). To calculate d, each experimental group was therefore contrasted with this baseline group on the immediate post-test.

4. For studies that did not involve control or comparison groups, but that reported pre-test and post-test values on a dependent variable (n = 5), effect size contrasts were drawn between the post-test and pre-test data for each experimental group (see discussion of this strategy in Light and Pillemer, 1984, p. 56).

5. A number of studies did not involve control or comparison groups and did not report pre-test dependent variable values, but did report post-test values for all experimental groups (n = 10). For these studies, one instructional treatment was selected as the baseline comparison condition, and d was calculated by contrasting each experimental group with this baseline condition on the immediate post-test. The baseline

condition was determined by identifying the least attention-focused treatment on a study-by-study basis.

It should be emphasized that strategies 3, 4, and 5 were adopted in order to include findings from as many studies, and about as many treatment types, as possible in the quantitative meta-analysis. On the basis of the contrasts described in (1)–(5) above, unique effect sizes were therefore calculated for all possible contrasts (i.e., for all independent variable conditions and all dependent variables) found within the 49 unique sample studies that provided sufficient data. Additionally, to investigate the extent of change from pre-experimental to post-experimental levels for all groups (including control/comparison groups), effect sizes were separately calculated for all studies reporting pre- and post-test values on dependent variables, in the manner described in (4) above.

Combining and comparing effect sizes. Combining effect sizes from individual studies enables the estimation of average effects related to instructional treatments. Prior to doing so in the current study, another issue related to the complexity of study designs within the research domain had to be resolved. It was noted in the previous section that effect sizes were calculated for all possible contrasts for a given unique sample study. Thus, a single study could produce multiple effect sizes, depending on how many independent and dependent variables were investigated (e.g., a study with two experimental conditions, a control condition, and three dependent variables would produce six effect size estimates). However, as the meta-analysis literature has pointed out (e.g., Cameron & Pierce, 1996), including multiple effect sizes from a single unique sample study leads to nonindependence of observations, and it also weights a given study more than other studies with fewer contrasts. In many meta-analyses, all contrasts from a single unique sample study are averaged, so that each study only contributes a single effect size estimate (e.g., Yirmiya, Erel, Shaked, & Solomonica-Levi, 1998). However, this strategy has been challenged for studies with comparisons among multiple

independent and/or dependent variables, as it may delimit or confound the actual range of findings reported in primary research (Wampold et al., 1997).

In the current meta-analysis, we adopted the following inclusive strategy. First, to represent the actual range of L2 instructional treatments investigated in primary research, effect sizes were calculated within each unique sample study by averaging for each independent variable all effect sizes resulting from different dependent variables (immediate post-tests). Thus, in most cases, each instructional treatment within a study was represented by a single effect size averaged across several dependent variable types. However, for any study that investigated the teaching of unrelated structures with the same instructional treatment, each structure taught was considered a unique independent variable with independent effect sizes. This is because there is evidence that structure may be a powerful moderating variable when assessing instructional effectiveness (see Alanen, 1995; de Graaff, 1997; DeKeyser, 1995; Ellis, 1993; Robinson, 1996; and discussion in Doughty & Williams, 1998, pp. 211–228). Although this approach resulted in the contribution of multiple effect sizes by several of the unique sample studies, and thus introduced non-independent values into the meta-analysis, it was adopted in order to provide the most representative picture of the instructional treatments that had received attention within the research domain.

For the purpose of summarizing findings related to the current research questions, effect sizes from unique sample studies were combined and compared in the following manner:

1. Average effect sizes were calculated for instructional treatment categories identified across studies, focusing specifically on FonF, FonFS, explicit, and implicit treatment types. In addition, average effect sizes were calculated for identifiable instructional treatment sub-types.

2. For all studies reporting pre-test levels on dependent variables, to investigate the amount of change observed within studies, average pre- to post-test effect sizes were calculated

for instructional treatments and for control/comparison groups. Zero prior knowledge studies were excluded, because they unnaturally inflate the effect size estimate (i.e., when the pre-test value is 0, it has no standard deviation and cannot be included in the calculation of *d*).

3. Average effect sizes were calculated on the basis of the duration of instructional treatment (according to four categories: brief, short, medium, long), in order to investigate a possible effect for briefer versus more extended instruction.

4. Average effect sizes were calculated for delayed post-tests, in order to investigate the durability of instructional effects over time.

5. Average effect sizes were calculated by type of dependent variable (according to four categories: constrained constructed response, free response, metalinguistic judgments, and selected response).

Of course, simply comparing averages only tells part of the cumulative story. It was also necessary to judge the statistical trustworthiness of effect size combinations (overall averages) and comparisons (between groups). Therefore, 95% confidence intervals were computed around each mean effect size by using a conservative random effects approach (Rosenthal, 1995, p. 187), according to the following formula, which treats the variability introduced by small sample sizes (adapted from Woods, Fletcher, & Hughes, 1986, pp. 102–103):

$$CI = d \pm \left[(95\%\ t = \text{distribution at } k - 1\ df) \left(\frac{sd}{\sqrt{k}} \right) \right] \qquad (4)$$

where *d* is the average effect size estimate, *sd* is the standard deviation of the average effect size estimate, and *k* is the number of study effect sizes.[14] As suggested by Rosenthal (1995), "The interpretation of this confidence interval is that if the claim of the effect size for the population [. . .] falls within the 95% confidence

interval, the claim will be correct 95% of the time" (p. 187). Although confidence intervals for effect sizes may also be calculated by using individuals within studies as the unit of analysis (as opposed to studies themselves), this approach is likely to produce a much narrower, and therefore overly optimistic, interval (Rosenthal, 1995).

Results

The Research Synthesis

To synthesize an overall picture of L2 type-of-instruction research as it is represented in the available literature, substantive, methodological, and reporting features were tallied and compared across the range of study reports. Patterns in research publication, as well as in the setting, design, and analysis of research studies, are summarized in this section, and summaries of substantive features are included with the quantitative meta-analysis in the following section.

Research publication. Experimental research on the effectiveness of L2 instructional treatments was published between 1980 and 1998 in 77 study reports that met the criteria for the current synthesis. As shown in Figure 1, publication within the research domain has seen a relatively steady increase since 1980, with a preponderance of the research (86%) published between 1990 and 1998. Study reports appeared in 21 different academic journals (including several institutional working papers) and in a number of edited books (see Appendix B). The majority of research within the domain was reported in three journals (38%), *Applied Linguistics*, *Modern Language Journal*, and *Studies in Second Language Acquisition*, or as chapters in edited collections (21%).

To investigate the likelihood of publication bias within the research domain (i.e., favoring the publication of studies that report statistically significant findings), the range of study effect sizes was graphed by the average group sample size within each

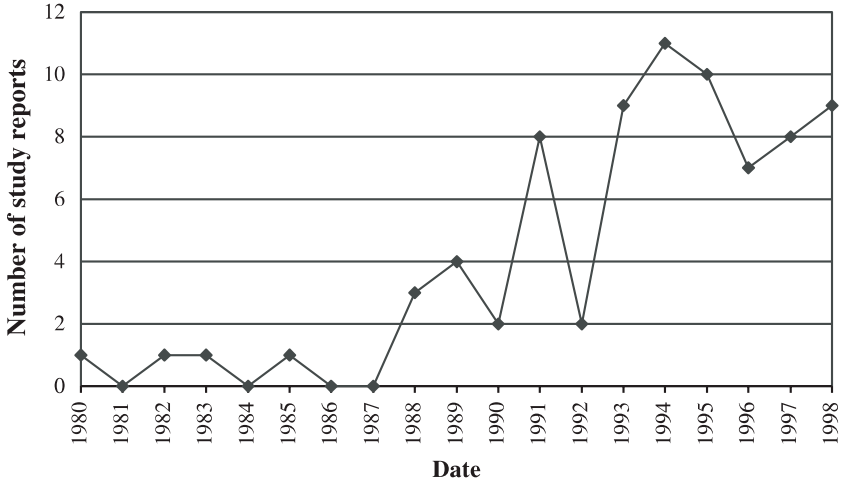


Figure 1. Publication frequency of studies reporting experimental research on L2 instructional treatments

study. In the absence of bias, one would expect to find a wide distribution of effect sizes associated with lower sample sizes, and an increasingly narrow distribution of effect sizes associated with larger sample sizes, in the approximate shape of an inverted funnel (Greenhouse & Iyengar, 1994; Light & Pillemer, 1984). That is, the lower the sample size, the larger the influence of sampling error, and the more diverse the expected observed effects. Thus, the influence of sampling variability should decrease as sample size increases, and observed study effect sizes should consistently cluster closer to the mean effect size. Individual effect sizes should also be distributed equally on either side of the mean (the center-line of the funnel), regardless of which sample size one looks at. In the presence of publication bias, one would expect to find "bites" taken out of the funnel shape, especially in the vicinity of effect sizes of zero, where statistically significant findings are less likely for smaller sample studies (Begg, 1994).

Figure 2 shows the distribution of effect sizes by sample sizes for studies in the current synthesis. Several patterns in this distribution can be noted. There is a definite clustering of effect

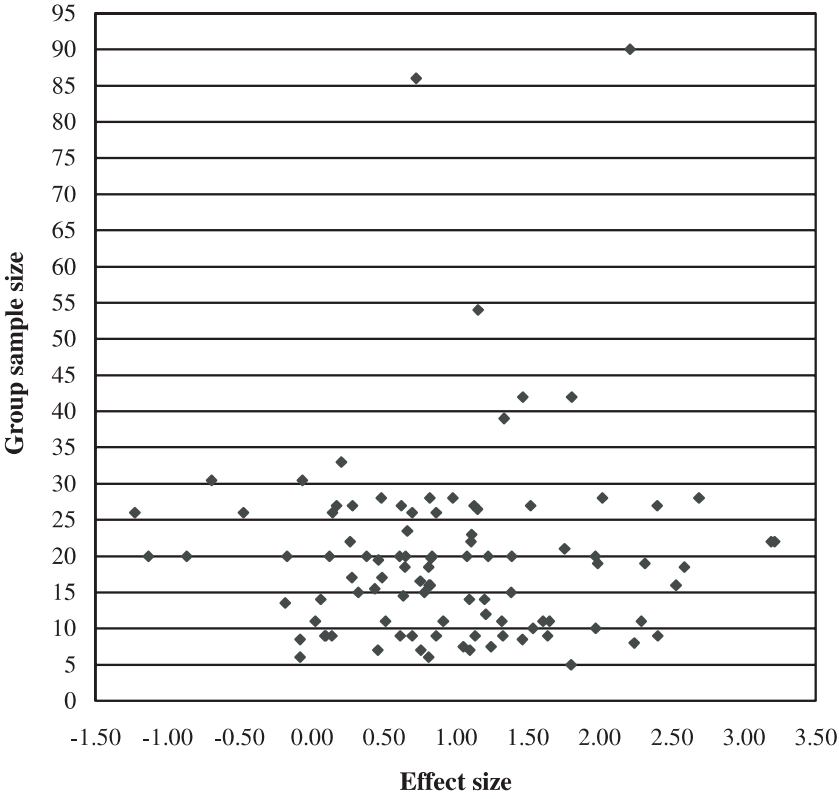


Figure 2. Effect sizes plotted against study group sample sizes for 78 unique sample studies (mean effect size, d = 0.96)

sizes around the mean effect size value (d = 0.96) at all sample sizes, and the large magnitude of this mean effect (showing on average one standard deviation of difference between experimental and control/comparison conditions) indicates the likelihood that most studies reported statistically significant findings, regardless of sample size. For group sample sizes of 20 or more, the funnel pattern seems to be consistent, with effect sizes distributed relatively equally on either side of the mean, and with decreasing dispersion of effect sizes as sample size increases. This indicates that studies with sample sizes of 20 or more have both larger and smaller effect sizes (and include a number of negative observed

effects as well). For studies with group sample sizes of fewer than 20, there is a marked truncation in the distribution of effect sizes, with "bites" taken out of the base of the funnel on both the positive and negative sides of the mean effect. Thus, small sample studies did not observe the range of large positive or negative effects one might expect as a result of sampling variability. However, some small sample studies did observe effects close to the zero value, indicating the likelihood that statistically nonsignificant findings were also reported in these studies. It should be noted here that many of the studies in the research domain investigated multiple instructional treatment conditions within a single investigation and are, therefore, represented several times within Figure 2. It is thus likely that, within a single study publication, effect sizes representing both statistically significant and statistically not significant findings were reported. The patterns observed in Figure 2 are, therefore, given further consideration in conjunction with the summary of statistical analyses below.

Research setting. L2 instructional treatments were investigated for several target languages, with a range of learner types, and in a number of instructional settings. Table 2 shows that eight different L2s served as targets for instruction, with English providing the target in 46% of the studies. Spanish, Japanese, and French in university FL contexts served as target L2s and instructional settings for 39% of the studies. Two languages, English and French, were investigated in diverse learning contexts: (a) as second languages (ESL, French Immersion); (b) as foreign languages (English/French FL); and (c) in immersion-like settings (intensive ESL programs in Canadian schools). Overall, 60% of the studies took place in foreign-language instructional settings, while the remaining 40% occurred in second language or immersion settings.

Table 2 also shows that learners came from a variety of L1s, although 50% of the studies were conducted with English L1 learners (i.e., virtually all of the non-English target FL studies were conducted with English L1 learners). A number of studies were conducted with mixed-L1 (15%) and Japanese L1 (17%)

learners. Investigations occurred within a range of educational contexts and with learners of differing ages, although studies were conducted largely with adult learners (79%) and in university settings (65%).

Learner proficiency levels within the target languages were reported inconsistently and, in general, minimally, with only a few studies reporting that initial learner proficiency levels had been established according to instruments such as TOEFL, ACTFL-type oral proficiency interviews, or in-house proficiency tests (Thomas, 1994, notes similar patterns in applied linguistics research in general). Virtually no studies established developmental levels of learners according to well-attested developmental sequences (although see exceptions in Doughty, 1991; Mackey & Philp, 1998; J. White, 1998). In general, researchers did not seem concerned about establishing with any rigor where learners may have been on a continuum from zero language ability to proficient

Table 2

Learner Characteristics in Research Domain

L2	*n*[a]	L1	*n*	Proficiency[b]	*n*	Educational Context	*n*
English SL	17	English	39	LOW	28	College	51
Spanish FL	16	Japanese	13	MID	16	Adult, not college	11
English FL	14	Mixed	12	HIGH	12	Junior high	10
French FL	8	French	7	MIX	6	High school	5
Japanese FL	7	Dutch	3	Not reported	16	Elementary	1
ESL intensive	5	German	1				
Artificial	5	Chinese	1				
French IM	4	Spanish	1				
Dutch SL	1	Not reported	1				
Finnish FL	1						
Welsh FL	1						

[a]*n* = number of unique sample studies reporting of $N = 78$ (the total in column 2 is 79, as a single study reported on results for two target L2s).

[b]LOW = learners with zero to two semesters of L2 study; MID = learners with three to four semesters of L2 study; HIGH = learners with five or more semesters of L2 study.

users of the L2. Furthermore, when researchers did report proficiency information, it was frequently inexact (e.g., "intermediate learners") and accompanied by no criteria for interpretation. However, a number of studies did report enough data to approximate the amount of L2 instruction that learners had received prior to the investigations, typically in the form of semesters or years of language study. On the basis of these data, learner populations within the research domain were separated into three rough proficiency levels: low, mid, and high (see definitions in Table 2). Table 2 shows that 36% of the studies investigated low-proficiency learners, 21% investigated mid-proficiency learners, and 15% investigated high-proficiency learners, while a handful of studies involved mixed-proficiency populations. Unfortunately, 21% of the studies reported no information about learner proficiency levels, so these patterns should be interpreted with caution.

Research design. The majority of L2 type-of-instruction studies (68%) adopted quasi-experimental designs, utilizing intact classes of learners for investigating treatment and control/comparison conditions. Several of these studies reported so-called random assignment of intact classes to different treatment levels. However, as no study involved more than a few classes, the statistical power and effectiveness of such random assignment was at best minimal (see Cook & Campbell, 1979, pp. 354–355). Furthermore, in virtually all of these cases, the unit of aggregation for statistical analyses of study findings was the individual study participant, and not the class.

A minority (31%) of studies, involving research volunteers, reported random assignment of individuals to treatment and control/comparison conditions. Study treatments were delivered in these studies in the following ways: (a) individually; (b) in small groups; (c) under laboratory settings; or (d) via computer-automated delivery systems. A single study (de Graaff, 1997) reported random sampling as well as random assignment of participants to research conditions.

Several other design features characterized studies within the research domain. Table 3 shows that 72% of the studies used

pre-tests to check for pre-experimental differences among research groups (unfortunately, 14% of these studies did not report the pre-test values). Of the 28% without a pre-test, eight studies (10% of the total sample) involved learners with zero initial knowledge of the target L2 (typically artificial languages), leaving only 18% of the studies having not attempted to establish the possibility of pre-experimental differences among learners. However, where some studies carefully measured or established pre-experimental learner ability levels on the structures being instructed, others reported pre-tests which measured constructs only weakly related to the instructional treatments or to post-tests. Control or non-focused (e.g., exposure-only) comparison groups were reported for 56% of the studies, although true control groups (i.e., no exposure to the targets) were operationalized in only 17% of the studies. The remaining 44% of the studies made direct comparisons among treatment conditions.

Table 4 indicates that sample sizes within the research domain varied widely, ranging from a low overall study sample of six participants to a high of 319 participants, although the mean (65.52) and mode (34) are more representative of the typical range of overall study sample sizes. Average group sample sizes tended to range between 5 and 35 participants (see Figure 2). The typical

Table 3

Study Designs in Research Domain

Design	n[a]
Pre-test	56
Control	13
Comparison	24
No control/comparison	19
No pre-test	22
Control	0
Comparison	7
No control/comparison	15

[a]n = number of unique sample studies reporting of N = 78.

instructional treatment ranged from less than an hour to around 4 hours, although some treatments involved as many as 50 hours of instruction. Virtually all studies reported post-experimental tests of treatment effectiveness, and these tests usually followed immediately or soon after the conclusion of instructional treatments. However, immediate post-tests were noted to follow as long as 26 days after instruction. Delayed post-tests were reported in 47% of the studies, and these follow-up measures usually occurred one to four weeks after instruction (although some occurred several months after instruction). A third post-test was reported in 18% of the studies, occurring between 2 and 48 weeks following instruction.

Finally, it should be pointed out that overall research designs ranged from very simple to quite complex, depending on the number of independent variables, moderator variables, and dependent variables investigated within a given unique sample study. Thus, simple study designs investigated as few as one treatment versus one control condition (or versus itself in repeated measures designs) on a single dependent variable. The most complex designs investigated as many as four to six independent variable conditions, as well as the influence of as many as four moderator variables, on up to four dependent variables. Moderator variables that were operationalized included: (a) learner aptitude; (b) learner awareness; (c) structural complexity; and (d) frequency of exposure to target-L2 tokens.

Table 4

Study Characteristics in Research Domain

Characteristic	Mean	*SD*	Range	Mode	*n*[a]
Sample size (*N*)	65.52	63.82	6–319	34	78
Length of treatment (*hours*)	4.08	7.01	.10–50	.25	75 (3 n.r.)
Timing of immediate post-test (*days*)	1.57	3.64	0–26	1	72 (6 n.r.)
Timing of delayed post-test 1 (*weeks*)	4.34	5.02	.71–24	1	37
Timing of delayed post-test 2 (*weeks*)	11.99	15.38	2–48	4	14

[a]*n* = number of unique sample studies reporting of *N* = 78; n.r. = not reported.

Research analysis. In analyzing their findings, researchers utilized a variety of observational, descriptive, and interpretive strategies. Several studies reported the use of qualitative techniques in analyzing the effects of instructional treatments, including: (a) think-aloud and retrospective protocol analysis; (b) observation and on-line coding of learner activities during instructional treatments; and (c) discourse analysis or related micro-analytic techniques. However, as the research domain was composed of experimental and quasi-experimental studies, virtually all researchers selected quantitative analyses as the primary means for describing and interpreting study findings. In general, both descriptive and inferential statistical techniques and data were reported inconsistently among the study reports, with the consequence that the analytic strategies, tools, and outcomes were often not sufficiently clear to enable readers to understand what was actually observed in the primary research.

Researchers described study findings using several standard statistical analyses. Table 5 shows the extent to which primary researchers reported descriptive analyses. All studies reported the overall number of study participants, although 18% of these studies did not report group sample sizes. Most studies reported some descriptive data about the dependent variable measures used, including the number and type of items tested, and a number of studies appended examples of such tests. However, where multiple forms of a test were utilized, very few studies reported information regarding how equivalency of these forms was established and whether or not the resulting measurement constructs were comparable. Furthermore, only 16% of the studies reported reliability estimates for the use of outcome measures.

Learner performance on dependent variables was reported in several ways. A few studies (14%) appended individual scores on the dependent variables for all research participants. More typically, studies reported average outcomes at the group level, with 82% of the studies presenting measures of central tendency on dependent variables. However, only 48% of the studies reported any measure of dispersion (e.g., standard deviations). Finally,

Table 5

Statistical Reporting (Descriptive)

Statistic	% studies reporting[a]
N_{overall}	100%
n_{group}	82%
Individual scores	14%
Mean (or other central tendency)	82%
S (standard deviation)	48%
Reliability (of outcome measures)	16%
Graphic	56%

[a]Based on $N = 77$ study report publications.

approximately half of the studies utilized graphic techniques for displaying research findings, occasionally in lieu of descriptive statistics.

Virtually all studies within the domain adopted statistical significance testing as the primary means for interpreting research findings. Table 6 shows the variety of analytic tools used as well as the extent to which the products of these tools were reported. It can be seen that 91% of the studies reported statistically significant findings (on at least one comparison among research groups), despite the fact that only 83% of the studies reported which statistical significance tests were used. One study reported "statistically significant" findings based on the use of no statistical significance tests whatsoever. The fact that virtually all of the studies reported at least one statistically significant comparison indicates a high probability for publication bias within the research domain.

A range of analytic tools were employed by primary researchers to conduct statistical significance comparisons among research groups, depending on complexity of research design or on researcher preference, and many of these techniques were used in conjunction with each other, including widespread use of multiple analyses of variance (ANOVAs) and multiple *t* tests with no corresponding adjustments in alpha levels and/or without the use

of preliminary multivariate analyses of variance (MANOVAs). Despite frequently low sample sizes, only limited use was made of non-parametric statistics (12% of studies). The third and fourth columns in Table 6 show that, although researchers reported using particular analyses, they did not always report the outcomes of these analyses in terms of the exact values of corresponding statistics. Only 26% of the studies displayed full results of such analyses in the form of inferential statistics tables.

The single most common approach to reporting and interpreting the results of statistical significance tests was by using probability levels. Of the studies in the research domain, 83%

Table 6

Statistical Reporting (Inferential)

Analysis component	% Employed[a]	Statistic	% Studies reporting[b]
MANOVA	11%	F	11%
ANCOVA	9%	F	7%
ANOVA	49%	F	45%
Post-hoc	43%	(exact)	20%
t test	33%	t	31%
Non-parametric	12%	(exact)	9%
Chi square	10%	χ^2	10%
Alpha (probability level)	83%	set $p < x$ a priori	20%
		exact p values	43%
		multiple p values	63%
Statistically significant finding	91%		
df (degrees of freedom)	62%		
Inferential statistics table	26%		
η^2 (strength of association)	5%		
Effect size	1%		
SE/CI (confidence interval)	3%		

[a]Based on $N = 77$ study report publications.

[b]Percentages of studies overall reporting the particular statistic produced by a particular analysis (e.g., 49% of the studies reported using ANOVA, but 4% fewer reported the resulting F value).

reported primary study findings according to whether or not comparisons were statistically significant at particular alpha levels. It is of interest to note that of these, only 23% (20% of all studies) established a priori acceptable probability levels to be applied across all significance tests. Many more studies (63% of all studies) reported that findings were statistically significant at multiple alpha levels, although only 43% of the studies reported exact *p* values for statistical significance tests employed. Finally, 8% of the studies reported statistically significant findings without setting or reporting any corresponding probability values.

By comparison with the use of statistical significance tests in interpreting study findings, only one study within the domain (Master, 1994) interpreted study findings by using an effect size index, although several other studies (5% of the study reports: Alanen, 1995; Doughty & Varela, 1998; Jourdenais et al., 1995; Leow, 1997) also utilized a strength of association index in addition to statistical significance tests. Finally, only 3% of the studies (Ellis et al., 1994; Robinson, 1996a) reported any measure of the trustworthiness of statistical comparisons, in the form of standard errors.

The Quantitative Meta-analysis of Substantive Study Findings

The quantitative meta-analysis focused on summarizing findings from available research about the effectiveness of different types of L2 instructional treatments by combining and comparing effect size estimates from individual studies. In addition, effect sizes were combined and compared by type of dependent variable, by duration of treatment, and for delayed post-tests. Overall, 49 unique sample studies contributed effect size estimates to these analyses. Given this relatively low number of studies, only general categories of interest could be summarized and interpreted with sufficient statistical power. Thus, although one might see the potential for comparing among studies on the basis of a wide variety of features, statistically trustworthy comparisons could only be made on the basis of variables

which are systematically represented across a large number of the studies.

For easier interpretation of the meta-analytic findings, all combinations of study effect sizes are presented according to a standard format in Tables 7 through 12. In addition to descriptive data, 95% confidence intervals are presented for each category within which study effect sizes were combined. These confidence intervals demonstrate the level of statistical trustworthiness with which average observed effects may be interpreted (Matt & Cook, 1994). The narrower the confidence interval, the more robust the observed effects. Furthermore, confidence intervals that do not include the zero value indicate that the observed effect differs probabilistically from the null hypothesis of no effect.

Instructional treatments. Within the research domain, L2 instruction has been operationalized as proceeding in terms of choices related to four components: presentation of rules, provision of negative feedback, exposure to relevant input, and opportunities for practice. Each of these four components presented multiple options for implementation, and any of the four elements could also be combined in various ways in a single instructional intervention, constituting particular pedagogical techniques (e.g., typographical input enhancement, input processing instruction, garden path). Within the research domain, researchers focused on some 20 different sub-types of L2 instructional treatments, but particular independent variables (sometimes with the same labels) differed from study to study. Compound types of instruction were also found that combined several components and techniques within a single instructional treatment, for instance, the so-called functional-analytic instruction investigated in the Canadian French immersion program studies (Day & Shapson, 1991; Harley, 1989, 1998; Lyster, 1994), the focus-on-form treatment in Leeman et al. (1995), and the explicit instruction treatment in Williams and Evans (1998). In sum, systematic replication of research on, as well as the accumulation of knowledge about, particular types of instruction is still incipient (this point is further addressed in the Discussion section).

Table 7 shows that, among those studies reporting data sufficient for calculation of effect sizes, an average of two types of instructional treatments plus a control/comparison/baseline condition were investigated within a single unique sample study (i.e., 49 unique sample studies contributed 98 effect sizes from unique independent variables). Researchers investigated from a minimum of one to a maximum of six independent instructional treatments within unique sample studies. Approximately 15% of the studies investigated four or more instructional treatments. Across all 49 studies, 56% of the instructional treatments were categorized as Focus on FormS, 80% of which involved explicit techniques. Of the 44% of treatments categorized as Focus on Form, 58% involved explicit techniques. Overall, 70% of the instructional treatments involved explicit techniques, while only 30% involved implicit techniques, and 40 unique sample studies operationalized at least one explicit treatment condition, while only 19 operationalized at least one implicit condition.

The effectiveness of L2 instruction. The average effect size observed across all instructional treatments (d = 0.96) indicates that treatment groups differed from control/comparison/baseline groups by approximately one standard deviation on immediate post-experimental outcome measures. Following Cohen's (1988) recommendation that effect sizes of 0.80 or greater should be considered large effects, this average overall effect size suggests that focused instructional treatments of whatever sort far surpass non- or minimally focused exposure to the L2 (the typical operationalization of baseline/comparison conditions). However, it should be noted that only 70% of the contrasts that produced this average overall effect size were based on differences between a treatment group and some sort of baseline or comparison condition (i.e., either exposure only or the least-focused instructional treatment), while 20% were based on true control conditions (i.e., receiving no exposure to the L2 target) and 10% were based on pre-to post-test differences in individual treatment groups. Furthermore, the high overall standard deviation (0.87) indicates that treatment effectiveness is quite widely dispersed around this mean

Table 7

Instructional Treatment Effect Sizes

IV	n^a	k^b	Mean *d*	*SD d*	95% CI lower	95% CI upper
Focus on FORM	25	43	1.00	0.75	0.78	1.22
Implicit	11	18	0.69	0.65	0.38	1.00
Flood		2	0.84	1.15	–9.45	11.13
Enhancement		5	0.56	0.71	–0.33	1.45
Recasts		4	0.81	0.78	–0.43	2.05
Other implicit		7	0.66	0.55	0.15	1.17
Explicit	18	25	1.22	0.75	0.91	1.53
Consciousness raising		3	1.78	0.55	0.36	3.20
Input processing		4	1.70	0.54	0.84	2.56
Compound FonF		10	1.00	0.50	0.64	1.36
Metalinguistic task-essentialness		3	1.68	1.34	–1.63	4.99
Rule-oriented FonF		4	0.55	0.66	–0.50	1.60
Focus on FORMS	32	55	0.93	0.96	0.67	1.19
Implicit	8	11	0.31	0.86	–0.27	0.89
Traditional implicit		1	–0.87	—	—	—
Corrective models		5	0.65	0.58	–0.07	1.37
Pre-emptive model		2	0.82	0.01	0.82	0.82
Form-experimental		3	–0.07	1.63	–5.02	4.88
Explicit	29	44	1.08	0.93	0.80	1.36
Traditional explicit		17	1.00	1.17	0.04	1.60
Input practice		3	1.87	0.44	0.79	2.95
Output practice		6	1.39	0.77	0.59	2.19
Rule-oriented forms-focused		6	0.57	0.32	0.24	0.90
Metalinguistic feedback		9	0.96	0.76	0.38	1.54
Garden path		3	1.50	1.04	–1.08	4.08
All implicit	19	29	0.54	0.74	0.26	0.82
All explicit	40	69	1.13	0.86	0.93	1.33
ALL TREATMENTS	49	98	0.96	0.87	0.78	1.14

[a]Number of unique sample studies contributing effect sizes.

[b]Number of instructional treatments contributing effect sizes (a single unique sample study could contribute multiple effect sizes when multiple instructional treatments were operationalized within the same study).

effect size. Observation error was nevertheless relatively small, given the overall substantial number of effect sizes contributed

by primary research, and the resulting 95% confidence interval relatively narrow (plus or minus 0.14 standard deviation units).

Mean combined effect sizes for particular categories of instructional treatment ranged on either side of this average effect size. On average, FonF treatments (d = 1.00) were observed to have slightly larger effect sizes than FonFS treatments (d = 0.93); and explicit treatments (d = 1.13) were observed to have substantially larger effect sizes than implicit treatments (d = 0.54). Based on average combined effect sizes for each category, the following pattern in instructional treatment effectiveness was observed among study findings:

FonF explicit > FonFS explicit > FonF implicit > FonFS implicit.

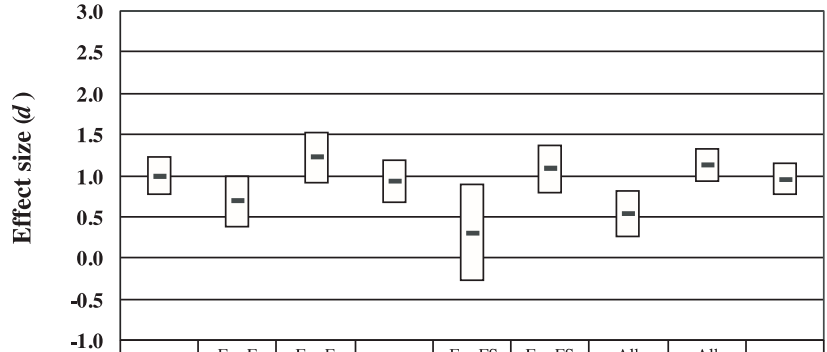
However, standard deviations were consistently high, indicating substantial heterogeneity among the effects observed within treatment categories. Although effect sizes for the two explicit categories would both be considered large effects according to the meta-analysis literature (Cohen, 1988), the average effect observed for FonF implicit treatments would only be considered a medium effect ($.50 < d < .80$), and that observed for FonFS implicit treatments a small effect ($.20 < d < .50$).

Confidence intervals of 95% displayed in the final two columns of Table 7 indicate that average observed effect sizes for particular treatment subtypes cannot be interpreted in any consistent or trustworthy way. Thus, confidence intervals are extremely broad for virtually all subtypes of study treatments, because each particular subtype has only been investigated in a handful of studies. Until more studies are conducted that systematically replicate each of these subtypes, comparisons should not be made among them (e.g., weighing the effectiveness of two particular treatments).

Confidence intervals are much narrower for the general categories of instructional treatments, and comparisons among them are therefore better warranted. Figure 3 displays the mean effect sizes and upper and lower 95% confidence boundaries for these categories. As shown in Figure 3, only one instructional

treatment category, FonFS implicit, includes the zero effect value within its 95% confidence interval, largely because only a few studies contributed data about implicit FonFS treatments. For all other categories, 95% confidence intervals fell well above the zero effect value, indicating that observed average effects for these groups differed probabilistically from no effect (this is equivalent to the interpretation that would be made with a test of statistical significance). Additionally, it should be noted in Figure 3 that average effect sizes for most of the treatment categories overlap with each other at the 95% confidence level. Thus, the observed differences in mean effectiveness between the different treatment categories fall within the realm of probabilistic sampling variability. In other words, although almost all of the instructional treatment categories differ substantially and with 95% probability from zero effects, observed differences between them may not be trustworthy. A single important exception can be noted in comparing the average overall effect sizes for explicit treatments versus implicit treatments, where 95% confidence intervals do not overlap. Thus, the observed difference in mean effectiveness between explicit and implicit treatments can be interpreted as a trustworthy difference.

To investigate the effectiveness of instructional treatments from another perspective, effect sizes were also calculated for the subset of studies reporting pre-experimental values on dependent variables (pre-tests). It will be recalled that this effect size calculation contrasted post-test values with pre-test values for all experimental groups within a given study, thus producing an estimate of the magnitude of change attributable to instructional treatments. Such pre- to post-test effect sizes were also calculated for all control or non-focused exposure comparison groups (but not for any least attention-focused treatments) in order to investigate the extent to which maturation or practice effect may be contributing to observed effects. Table 8 shows that 19 unique sample studies reported sufficient data for calculating pre- to post-test effect sizes and that 14 of these also reported data on true control/comparison conditions.



	FonF (*k*=43)	FonF implicit (*k*=18)	FonF explicit (*k*=25)	FonFS (*k*=55)	FonFS implicit (*k*=11)	FonFS explicit (*k*=44)	All implicit (*k*=29)	All explicit (*k*=69)	Overall (*k*=98)
CI upper (95%)	1.22	1.00	1.53	1.19	0.89	1.36	0.82	1.33	1.14
Mean	1.00	0.69	1.22	0.93	0.31	1.08	0.54	1.13	0.96
CI lower (95%)	0.78	0.38	0.91	0.67	-0.27	0.80	0.26	0.93	0.78

Figure 3. Average effects and 95% confidence intervals for instructional treatment categories

On average, instructional treatments induced 1.66 standard deviations of change in performance from pre-test values to post-test values on study outcome measures. Once again, it can be observed that FonF treatments were associated with greater change than were FonFS treatments, although low numbers of studies reporting sufficient data render comparisons among explicit and implicit conditions within each of these categories too unstable for trustworthy interpretation. It should also be noted that control or comparison conditions exhibited change (d = 0.30) from pre-test levels to post-test levels as well (of these conditions, 43% involved some kind of nonfocused exposure to the L2, while 57% were true control conditions). This observation concurs with a similar pattern in control group change observed by Hulstijn (1997) among laboratory-based L2 instructional studies. Furthermore, whereas the large standard deviations for change induced by instructional treatments indicate that such pre- to post-test change varied widely from study to study, much lower standard deviation values were noted for control/comparison conditions,

Table 8

Magnitude of Change From Pre-test to Post-test

IV	n^a	k^b	Mean *d*	*SD d*	95% CI lower	95% CI upper
Treatment	19	43	1.66	0.95	1.36	1.96
Focus on FORM	12	18	1.92	1.01	1.42	2.42
Implicit	4	5	1.51	0.91	0.37	2.65
Explicit	11	13	2.08	1.03	1.45	2.71
Focus on FORMS	13	25	1.47	0.88	1.10	1.84
Implicit	2	2	1.87	1.75	–13.88	17.62
Explicit	12	23	1.43	0.83	1.08	1.78
Control/comparison	14	15	0.30	0.39	0.19	0.41

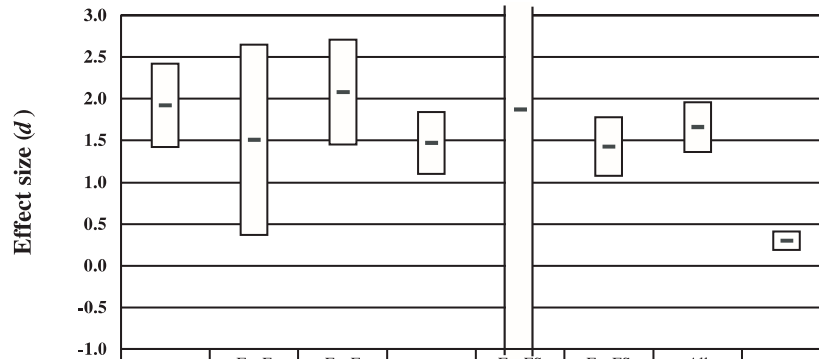
[a]Number of unique sample studies contributing effect sizes.
[b]Number of instructional treatments contributing effect sizes (a single unique sample study could contribute multiple effect sizes when multiple instructional treatments were operationalized within the same study).

indicating that change was much more consistent among these groups. It can be inferred from these observations that, for treatment conditions, as much as 18% of change from pre- to post-test levels on dependent variables is attributable to something (e.g., practice effect, exposure-only effect, maturation) besides the effect of a given instructional treatment.

Figure 4 shows the magnitude of change from pre-test to post-test as well as the 95% confidence intervals associated with these observations. Once again, it should be noted that confidence intervals for instructional treatment categories do not include zero values, with the exception of the FonFS implicit category, where the massive confidence interval can be attributed to the contribution of effect sizes from only two studies with rather divergent findings. In addition, however, intervals are relatively broad and overlap with each other, indicating that observed mean differences among treatment types may not be trustworthy. Finally, it should be noted that the 95% confidence interval around the mean change observed for control/comparison conditions is quite narrow (plus or minus 0.11 standard deviation units). It can therefore be

inferred with some confidence that control/comparison groups within studies in the research domain will exhibit consistent change towards the target of instruction over the course of a study.

Magnitude of effect for outcome measures. Within L2 type-of-instruction research, researchers employed a variety of different outcome measures as dependent variables to test the effectiveness of instructional treatments. Such measures ranged from discrete point tests, which asked research participants to display grammatical knowledge, to free oral production, which was later coded and analyzed by researchers. Overall, interpretations based on the constructs represented within these outcome measures were of three types: (a) Did participants acquire the ability to recognize the target form? (b) Did participants acquire the ability to produce the target form? (c) Did participants acquire the ability to explain the rule-governed nature of the target form? Length and difficulty of outcome measures varied depending on the targeted structures, the learners, the amount of instruction, institutional factors (e.g.,



	FonF (*k*=18)	FonF implicit (*k*=5)	FonF explicit (*k*=13)	FonFS (*k*=25)	FonFS implicit (*k*=2)	FonFS explicit (*k*=23)	All treatments (*k*=43)	Control (k= 15)
CI upper (95%)	2.42	2.65	2.71	1.84	17.62	1.78	1.96	0.41
Mean	1.92	1.51	2.08	1.47	1.87	1.43	1.66	0.30
CI lower (95%)	1.42	0.37	1.45	1.10	-13.88	1.08	1.36	0.19

Figure 4. Average magnitude of change from pre-test to post-test for instructional treatment and control conditions

time allotted for the research), researcher preference, and a number of factors idiosyncratic to individual studies.

Studies varied in the extent to which targeted formal aspects of the L2 were tested on outcome measures, with some studies utilizing as little as one test item per targeted structure to inform interpretations about the effectiveness of instructional treatments. Other studies utilized lengthy tests with multiple items per targeted structure or collected extensive L2 production data. Performance on outcome measures was evaluated by primary researchers in the following ways: (a) according to dichotomous criteria (correct or incorrect); (b) according to polytomous criteria (e.g., subjective ratings); (c) according to interlanguage sensitive criteria (weighting L2 production according to various target-oriented stages); or (d) according to suppliance or error frequency counts. On average, individual studies utilized between two and three dependent variables (with a minimum of one and a maximum of four in any single study) to ascertain the effectiveness of instructional treatments.

For the 49 unique sample studies reporting sufficient data, average effect sizes were calculated for four categories of dependent variables, on the basis of the types of responses required from research participants. Table 9 summarizes findings related to these categories. It should be noted that the majority of the studies (65%) employed constrained constructed response measures, while fewer studies utilized outcome measures with other response types (39% used selected response, 29% used metalinguistic judgment, and 16% used free constructed response).

Table 9 shows that average effect sizes associated with metalinguistic judgments and free constructed response measures were substantially lower than those associated with selected-response or constrained constructed-response measures. Thus, study findings within the research domain may vary by as much as 0.91 standard deviation units depending on the type of outcome measure or measures employed. Standard deviations within each of the four categories also reflect a large degree of variability among findings from studies utilizing the same type of outcome

Table 9

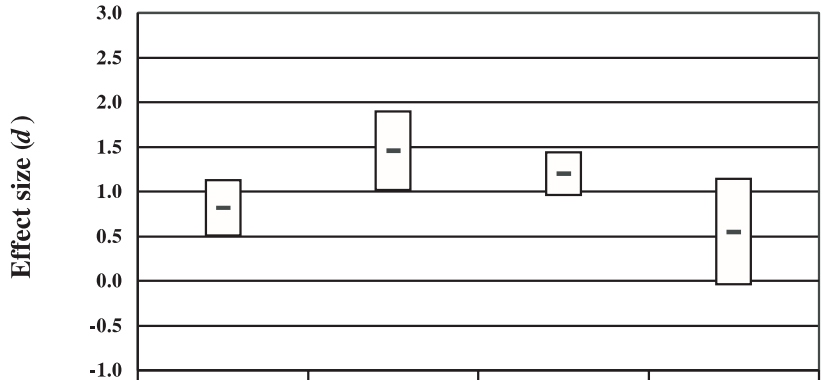
Magnitude of Effect by Types of Outcome Measure

DV	n^a	k^b	Mean *d*	*SD d*	95% CI lower	95% CI upper
Meta-linguistic judgment	14	29	0.82	0.79	0.51	1.13
Selected response	19	32	1.46	1.23	1.02	1.90
Constrained constructed response	32	62	1.20	0.95	0.96	1.44
Free constructed response	8	13	0.55	0.97	–0.04	1.14

[a]Number of unique sample studies contributing effect sizes.
[b]Number of instructional treatments contributing effect sizes (a single unique sample study could contribute multiple effect sizes when multiple instructional treatments were operationalized within the same study).

measure. Figure 5 displays the differences among mean effect sizes associated with the four types of outcome measures, although 95% confidence intervals also overlap for all four types. Therefore, although substantial mean differences seem to suggest variability in effect due to type of outcome measure employed, these observations may not be beyond the realm of probability.

Given the substantial observed differences among types of outcome measures, an association between particular instructional treatment categories and particular outcome measure categories could account for differences observed in the effectiveness of different L2 treatment types. To investigate this possibility, the percentage of outcome measures utilized within each of the instructional treatment categories was calculated. Figure 6 shows the percentage distributions of outcome measure types by instructional treatment types. Within the FonF category, distributions were very similar, although some degree of difference may be attributed to the fact that FonF explicit treatments utilized more selected response measures (d = 1.46) and fewer metalinguistic judgment measures (d = 0.82) than did FonF implicit treatments. Within the FonFS category, marked differences can be observed between explicit and implicit treatments. Thus, explicit treatments used many more selected response measures (d = 1.46) and



	Meta-linguistic judgment (*k*=29)	Selected response (*k*=32)	Constrained constructed response (*k*=62)	Free constructed response (*k*=13)
CI upper (95%)	1.13	1.90	1.44	1.14
Mean	0.82	1.46	1.20	0.55
CI lower (95%)	0.51	1.02	0.96	-0.04

Figure 5. Average effects and 95% confidence intervals for types of outcome measures

fewer free constructed response measures (d = 0.55) than did implicit treatments. However, these differences are ameliorated by the fact that implicit treatments used many more constrained constructed-response measures (d = 1.20). Finally, it should be noted that a range of outcome measures was utilized within each of the four instructional treatment types. In general, there does not seem to be a pattern of instructional treatments with lower mean effect sizes having predominantly utilized outcome measures with lower mean effect sizes. Thus, although some of the differences in effects observed among treatment categories may be attributable to differential use of outcome measures, it is unlikely that the magnitude of observed differences can be accounted for in this way.

Duration of treatment and durability of effect. To investigate whether amount of instruction influenced the effectiveness of instruction, effect sizes were combined for studies that investigated treatments of similar duration. Four general categories of

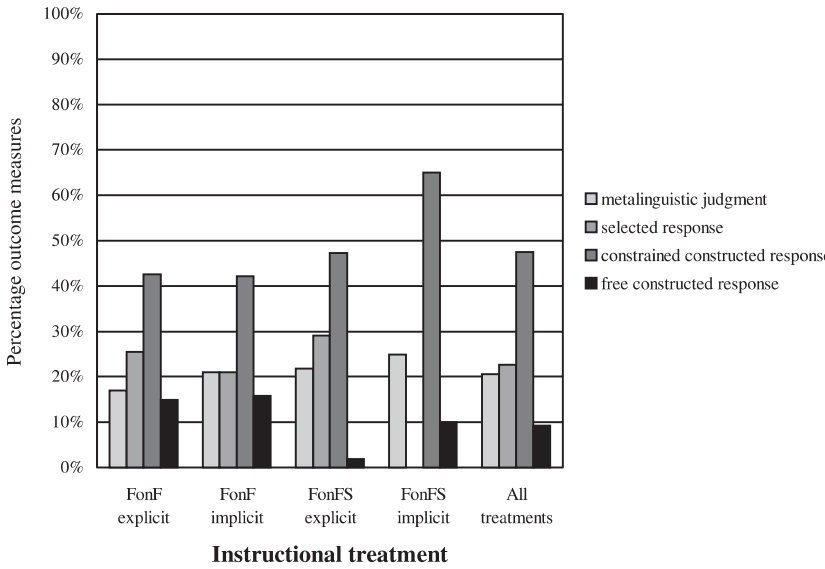


Figure 6. Percentage of outcome measure types by instructional treatment categories

treatment duration were identified within the research domain (see definitions in Table 10), and average effect sizes were calculated for each of these categories. Table 10 shows that 68% of instructional treatments lasted for less than two hours, while the remaining 32% lasted for three hours or longer. Virtually no differences in mean effect sizes were observed within further subdivisions of either of these categories. Treatments of less than one hour differed on average by only .02 standard deviations from treatments of one to two hours, and treatments of three to six hours did not differ on average from treatments in excess of seven hours. However, once again, standard deviations within each of these subdivisions were quite large, indicating substantial variation in effect sizes among individual studies.

The average effect size observed for treatments of less than two hours (d = 1.07) did differ substantially (by nearly one third of a standard deviation unit) from the average effect size observed for treatments of three or more hours (d = .79). Thus, shorter-term

Table 10

Magnitude of Effect by Duration of Treatment

Amount of instruction	n[a]	k[b]	Mean d	SD d	95% CI lower	95% CI upper
Brief treatment (x < 1 hr)	16	34	1.06	1.02	0.71	1.41
Short treatment (1 hr < x < 2 hr)	14	33	1.08	0.79	0.79	1.37
Medium treatment (3 hr < x < 6 hr)	10	17	0.79	0.89	0.32	1.26
Long treatment (x > 7 hr)	9	14	0.79	0.94	0.24	1.34

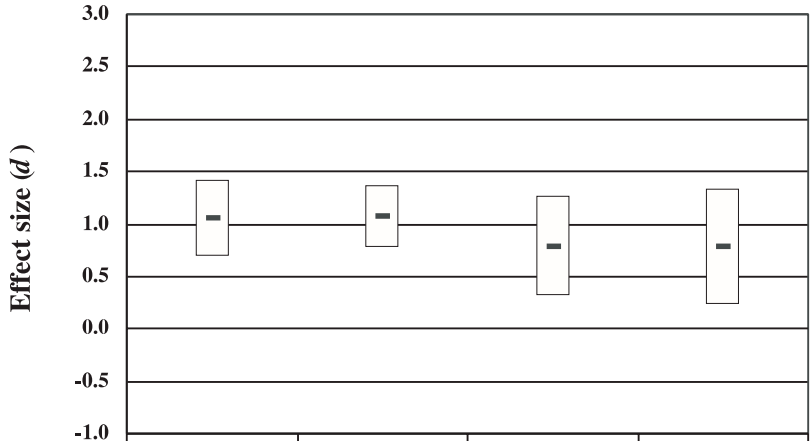
[a]Number of unique sample studies contributing effect sizes.
[b]Number of instructional treatments contributing effect sizes (a single unique sample study could contribute multiple effect sizes when multiple instructional treatments were operationalized within the same study).

instructional treatments seem to produce larger effects than longer-term treatments. Of course, it should not be assumed that this is a causal relationship, suggesting that less instruction is more effective. Indeed, it is likely that a combination of moderating factors accounts for these observed differences, including: (a) the types of structures instructed within shorter- versus longer-term treatments; (b) immediacy and construct proximity of outcome measures; and (c) the relative intensity of instruction within shorter- versus longer-term treatments, among others. Unfortunately, none of these potential moderator variables has been operationalized to date in a consistent way across the full range of studies. Attempting to construct explanatory models based on such variables would, therefore, be a premature undertaking, given the present state of research and reporting within the domain.

Figure 7 shows the mean effect sizes and associated 95% confidence intervals for each category of treatment duration. It should be noted that confidence intervals are broader for the longer treatments, because fewer studies have been conducted on longer-term instruction. Once again, although the mean differences between shorter- and longer-term treatments are obvious in

Figure 7, these differences should be noted to overlap at the 95% confidence level. Observed differences between treatments of different durations therefore fall within the realm of probabilistic variability.

Given the observed differences in mean effects associated with shorter- versus longer-term treatments, an association between length of treatment and instructional treatment categories could account for differences observed in the effectiveness of different L2 treatment types. Thus, for example, a preponderance of longer-term treatments within a particular instructional treatment category might account for lower observed effect sizes. Figure 8 shows for each instructional treatment category the corresponding percentages of effect sizes contributed by treatments of differing lengths. No patterns can be noted among the instructional treatment categories with lower effect sizes (FonF implicit and FonFS implicit) and the duration of treatment categories with lower effect sizes (medium- and long-term treatments).



	Brief treatment (*k*=34)	Short treatment (*k*=33)	Medium treatment (*k*=17)	Long treatment (*k*=14)
CI upper (95%)	1.41	1.37	1.26	1.34
Mean	1.06	1.08	0.79	0.79
CI lower (95%)	0.71	0.79	0.32	0.24

Figure 7. Average effects and 95% confidence intervals for differing lengths of instructional treatments

In fact, within the FonF treatment categories, the implicit type consisted of approximately 15% more brief and short-term treatments (d = 1.07) than did the explicit type, even though the explicit-type treatments showed overall higher average effects (see Table 7). Likewise, within the FonFS treatment categories, implicit types consisted of more brief and short-term treatments (91%) than medium- and long-term treatments (9%), and approximately 20% more shorter-term treatments overall than for the explicit types, even though FonFS implicit treatments showed much lower average effect sizes.

A number of studies within the research domain investigated the durability of instructional treatment effectiveness, typically in the form of one or more delayed post-tests. Table 11 shows that 22 unique sample studies reported data sufficient for calculating effect sizes on an immediate and a single delayed post-test. Across all treatment types, observed effectiveness of

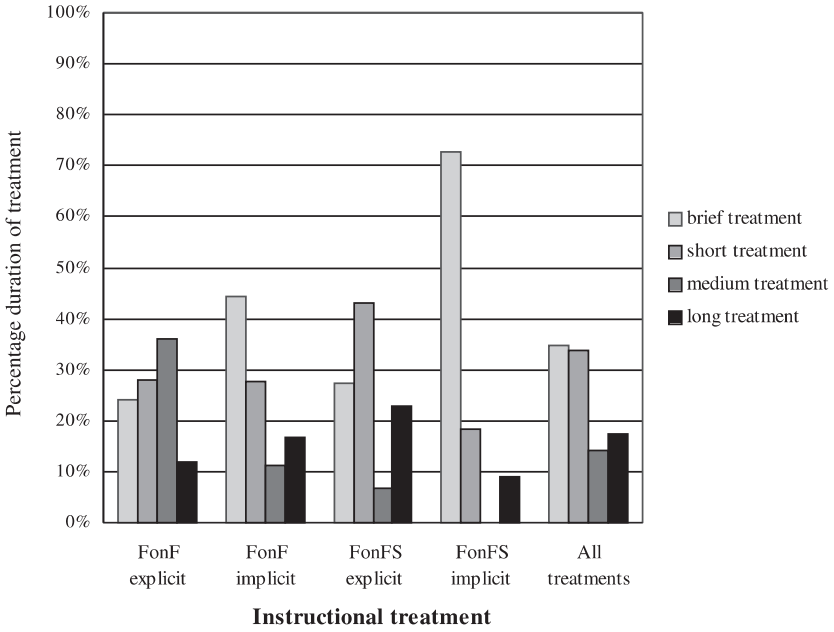


Figure 8. Percentage of effect sizes contributed by studies with differing lengths of treatment

instructional treatments was generally maintained, although the observed effect was reduced on average by one-fifth of a standard deviation unit from the immediate post-test to the delayed post-test. This reduction in the observed effect size was also in evidence across treatments involving differing amounts of instruction, although the numbers of studies contributing effect sizes to these averages were too few to foster trustworthy comparisons among instructional treatments of differing lengths (see 95% confidence intervals in Table 11).

Table 11 also shows that the overall pattern of durable but truncated effect sizes was accompanied across all treatment types and durations by a reduction in standard deviations. Thus, the standard deviation is always smaller on the delayed post-test effect size than on the immediate post-test effect size. Given the

Table 11

Durability of Effect for One Delayed Post-test

Amount of instruction	n[a]	k[b]	Mean d	SD d	95% CI lower	95% CI upper
Brief treatment (x < 1 hr)						
Immediate post-test	5	11	1.35	1.28	0.50	2.20
Delayed post-test	5	11	0.93	1.06	0.22	1.64
Short treatment (1 hr < x < 2 hr)						
Immediate post-test	9	20	1.42	0.78	1.06	1.78
Delayed post-test	9	20	1.28	0.67	0.97	1.59
Medium treatment (3 hr < x < 6 hr)						
Immediate post-test	4	4	0.73	0.34	0.26	1.20
Delayed post-test	4	4	0.69	0.33	0.22	1.16
Long treatment (x > 7 hr)						
Immediate post-test	4	7	0.84	0.52	0.35	1.33
Delayed post-test	4	7	0.61	0.46	0.19	1.03
ALL TREATMENTS						
Immediate post-test	22	42	1.24	0.89	0.96	1.52
Delayed post-test	22	42	1.02	0.77	0.78	1.26

[a]Number of unique sample studies contributing effect sizes.

[b]Number of instructional treatments contributing effect sizes (a single unique sample study could contribute multiple effect sizes when multiple instructional treatments were operationalized within the same study).

change in control group behavior from pre-test to post-test noted in Table 8 (q.v.), it is likely that the smaller standard deviations on delayed post-test effect sizes are attributable not only to some loss of instructional effect on the part of treatment groups but also to some amount of continued maturation by control or comparison groups. However, it should be noted that observed advantages for instruction tend to remain robust across delayed post-test comparisons (note in Tables 11 and 12 that most 95% confidence intervals do not include the zero value).

Table 12 shows that only six unique sample studies contributed sufficient data for calculating effect sizes on an immediate and two delayed post-tests. It is interesting to note that these studies only operationalized brief or short-term treatments. Overall, for this set of studies, decreasing effect sizes can be noted on average from the immediate to the delayed to the second delayed post-tests. However, the overall decrease from the immediate post-test to the second delayed post-test is on average less than three-tenths of a standard deviation unit. Therefore, effects seem to be relatively durable. Consistent decreases in standard deviations can also be noted from immediate through delayed post-tests, and this may indicate once again both a loss of instructional effect on the part of treatment groups and some degree of maturation on the part of control or comparison groups. Comparisons of the durability of effects between treatments of differing durations are not trustworthy, as the number of studies contributing effect sizes is too few.

Discussion

Research Question 1: How Effective Is L2 Instruction Overall and Relative to Simple Exposure or Communication?

Robey (1998) has noted that "[t]he products of a meta-analysis, the average effect size and its confidence interval, estimate the degree to which a particular null hypothesis is false on the basis

Table 12

Durability of Effect for Two Delayed Post-tests

Amount of instruction	n^a	k^b	Mean d	SD d	95% CI lower	95% CI upper
Brief treatment (x < 1 hr)						
Immediate post-test	3	7	1.45	1.59	–0.02	2.92
Delayed post-test 1	3	7	1.07	1.35	–0.18	2.32
Delayed post-test 2	3	7	1.00	1.19	–0.10	2.10
Short treatment (1 hr < x < 2 hr)						
Immediate post-test	3	5	1.14	0.99	–0.08	2.36
Delayed post-test 1	3	5	1.09	0.67	0.26	1.92
Delayed post-test 2	3	5	1.16	0.54	0.49	1.83
ALL TREATMENTS						
Immediate post-test	6	12	1.32	1.33	0.49	2.15
Delayed post-test 1	6	12	1.17	1.18	0.43	1.91
Delayed post-test 2	6	12	1.06	0.94	0.47	1.65

[a]Number of unique sample studies contributing effect sizes.

of all available evidence" (p. 173). With respect to RQ 1, the null hypothesis would posit no difference in effectiveness between instructional treatments and control/comparison or baseline treatments as measured on dependent variables.

An absolute (nonrelative) estimate of the effectiveness of L2 instructional treatments was investigated by contrasting post-test values for individual treatments with the pre-test values for those treatments (see results in Table 8). On average, L2 instructional treatments induced 1.66 standard deviations of target-oriented change in L2 ability or knowledge. Although findings across studies were heterogeneous, the substantial magnitude of this average pre- to post-treatment change (the lower 95% confidence interval boundary of which was 1.36 standard deviation units above the level of zero effect) suggests that instructional treatments are quite effective, and that observed effects are quite different from the null hypothesis of no difference. This finding

must be qualified by the fact that, on average, 18% of this amount of pre- to post-test change was also noted for true control/comparison groups. Nevertheless, L2 instruction can be characterized as effective in its own right, at least as operationalized and measured within the domain. The average overall magnitude of change noted in the current synthesis should serve as a useful index for interpreting the pre- to post-test effectiveness of L2 instructional treatments (and change in control/comparison conditions) in future investigations.

Further evidence for the overall effectiveness of L2 instruction was also sought in the main body of study effect sizes, by contrasting treatment groups versus control/comparison or baseline groups on immediate post-test values (see results in Table 7). Effect sizes aggregated across 49 unique sample studies indicated that focused L2 instructional treatments consistently outperformed a range of control/comparison or baseline conditions by an average of nearly one standard deviation unit (d = 0.96), by all accounts a large and convincing magnitude of effect (Cohen, 1988). Owing to the number of studies investigating this issue, and to the consistent magnitude of difference between focused instruction versus control/comparison and baseline conditions, this average finding was also noted to be quite trustworthy, with a relatively narrow 95% confidence interval (plus or minus 0.18 standard deviation units), the lower boundary of which differed positively from zero (the null hypothesis of no difference) by 0.78 standard deviation units. In short, not only does focused L2 instruction make a consistently observable difference that is very unlikely to be attributable to chance, but it also seems to make a substantial difference.

The question of just how effective instruction is when compared with simple exposure and/or simple communication is not directly answered by the comparisons shown in Table 7. It should be recalled that the average effect sizes there were calculated on the basis of accumulated findings from: (a) the 70% of the 49 unique sample studies that operationalized a baseline or comparison condition (i.e., some kind of a least-focused instruction

condition or exposure-only group); (b) the 20% of the studies that contributed effect sizes based on contrasts between treatment and true control conditions (where the only control group exposure to the L2 targets occurred in pre- and post-test sessions); and (c) the additional 10% of the studies that contributed effect sizes based only on pre- to post-test changes in treatment groups (where no control/comparison or baseline conditions were operationalized). Recalculation of an average effect size for the 70% of studies whose designs compared focused versus nonfocused comparison groups or least attention-focused instructional baseline treatments offers an estimate of the magnitude of effect for (focused) instructional treatments when compared with simple exposure to the L2 targets, experience with the L2 tasks, or some minimal amount of both. On average, a substantial effect was still observed (d = 0.75, S = 0.85), although smaller than that when contrasts with true controls were included for average effect size calculations. The 95% confidence interval produced upper (1.06) and lower (0.64) boundaries for this average effect that again demonstrated substantial difference from the null hypothesis. This mean effect size (approximately three fourths of a standard deviation unit) thus provides some evidence of the extent to which focused L2 instructional treatments surpass nonfocused treatments in terms of effectiveness. However, this finding must be qualified by noting that such direct contrasts may not reveal the true difference in effectiveness between such instructional conditions (see discussion of RQ 6).

Research Question 2: What Is the Relative Effectiveness of Different Types and Categories of L2 Instruction?

While the wide variety of instructional treatment types investigated within the domain renders more robust the finding of a large average effect for focused instruction, it also reduces the likelihood of finding consistencies among particular instructional treatment types. That is, no particular sub-types of L2 instructional delivery have been the subject of systematic replication

sufficient for drawing cumulative inferences about their relative effectiveness. At a general level, however, studies have with more consistency investigated instructional treatments that were reliably categorized according to whether or not there was an integration of form and meaning (FonF versus FonFS instruction) and whether or not rule explanation or related attention to the rule-governed nature of L2 structures was incorporated into the treatment (explicit versus implicit instruction).

Both FonF and FonFS instructional categories were observed to have large average effect sizes of around one standard deviation unit, and the two categories differed by only 0.07 standard deviation units from each other. Given the proximity of these average effects, as well as wide variation in individual study effect sizes within each of these groups, 95% confidence intervals were largely overlapping. Thus, although each category of instructional treatment differed substantially from the null hypothesis level (with lower boundaries 0.78 standard deviation units above the zero effect for FonF and 0.67 standard deviation units above the zero effect for FonFS), observed differences between the two categories are not trustworthy. Furthermore, assuming that future studies would reproduce the distribution of effect sizes observed for each category, it is unlikely that adding any number of studies would narrow the confidence intervals around these average effect sizes to such an extent that trustworthy differences would be observed between FonF and FonFS treatments. Thus, current cumulative research findings suggest no differences in effectiveness between FonF and FonFS instruction (as currently operationalized) and equivalent overall instructional effectiveness for both.

The trend for explicit versus implicit treatments is different from that observed for FonF versus FonFS treatments. The average observed effect for explicit treatments (d = 1.13) differed by more than half a standard deviation unit from the average effect for implicit treatments (d = 0.54), and 95% confidence intervals around these two observed effect sizes did not overlap, indicating a trustworthy observed difference. Thus, the current state of findings within this research domain suggests that treatments

involving an explicit focus on the rule-governed nature of L2 structures are more effective than treatments that do not include such a focus.

To carry out a more fine-grained analysis of relative instructional effectiveness, the potential interaction of FonF/FonFS and explicit/implicit categories was also investigated, and interpretable trends in effectiveness were noted. Thus, within FonF instruction, implicit treatments were noted on average to be half a standard deviation unit less effective than explicit treatments. Although their 95% confidence intervals overlapped, indicating that observed differences may be due to sampling error, adding another 10 studies in each category would result in nonoverlapping 95% confidence intervals if the distribution of observed effects remained unchanged (thus rejecting the null hypothesis of no difference between the categories). Likewise, within FonFS instruction, a large mean advantage for explicit over implicit categories was noted (0.77 standard deviation units). Adding an additional four studies in each of these categories would result in nonoverlapping confidence intervals (and rejection of the null hypothesis), if the distribution of observed effects remained unchanged.

There are a number of possible explanations for differences or lack of difference observed among FonF/FonFS and explicit/implicit categories of L2 instructional treatments. As many have pointed out (e.g., Robinson, 1996a; Tomlin & Villa, 1994), the measurement of change induced by instruction is typically carried out on instruments that seem to favor more explicit types of treatments by calling on explicit memory-based performance. Thus, in the current domain, over 90% of the dependent variables required the application of L2 rules in highly focused and discrete ways, while only around 10% of the dependent variables required relatively free productive use of the L2 (see discussion below). In addition, most primary research has operationalized implicit treatments in relatively restricted ways, whereas explicit treatments often involve combinations of several instructional components. Thus, a typical explicit treatment may include rule

presentation, focused practice, negative feedback, and rule review, whereas an implicit treatment may simply involve a single type of implicit exposure.

It should also be recalled that heterogeneity in effects was observed within all instructional treatment categories. Such heterogeneity likely occurs in large part because individual studies operationalize instructional treatments via widely differing independent variables and because variables are not consistently replicated from study to study. Two distinct associated problems were observed repeatedly across primary studies in the current synthesis. First, various features of an instructional component were sometimes merged in a single instructional intervention without precise control (or description of) such features as they occurred during treatment delivery. Second, treatments that were intended to be an operationalization of the same instructional type did indeed vary from study to study.

The first problem can be illustrated with the element of rule explanation. Presentation of rules in most explicit treatments was paradigmatic, with various forms and functions of a linguistic subsystem presented together. However, rule presentation in input processing treatments was staged (Cadierno, 1992), with aspects of a structure explained in small steps accompanied by intervening practice or exposure activities (e.g., Cadierno, 1992, 1995; and VanPatten & Cadierno, 1993a, 1993b). In addition, most rule-based treatments delivered grammar explanations a priori, before engaging in other types of instructional activities, whereas grammar explanations were made available to learners for consultation throughout the instructional activities in Robinson's (1996b, 1997) instructed group, and they were repeated over the course of intervention at certain intervals in DeKeyser's (1997) explicit condition. Whether rule explanation is paradigmatic or staged, presented once or repeated, and available for memory scaffolding throughout the treatment or not, could make a difference in the observed effectiveness of the specific instructional types (e.g., Leow, 1998a). However, such variations are rarely

described in detail, let alone controlled for or systematically operationalized as moderator variables from study to study.

The second problem, that of substantive differences in operationalizations of purportedly the same instructional type, can be illustrated with the provision of negative feedback in various treatments. It was often the case that a given type of feedback was delivered in one study through several feedback moves compounded within a treatment and in another study within a single feedback move. One case of a "simple" versus "compounded" version of the same instructional feedback type was found in clarification request treatments. In Nobuyoshi and Ellis (1993), clarification requests involved a single move and a pause for learner self-correction after a nontargetlike utterance, whereas in Herron and Tomasello (1988), the treatment involved a cycle of clarification requests and opportunities for self-correction after a single nontargetlike utterance.[15] Similarly, recasts present another case of simple versus compounded variations of a single instructional type. Thus, the recast treatment in Doughty and Varela (1998) sometimes involved repetition of a learner's error with intonational enhancement and opportunity for the learner to reformulate both before and after the targetlike form was provided by the teacher, whereas these additional features were absent in the recast treatments in Long, Inagaki, and Ortega (1998) and in the intensive regime of recasts delivered by trained native speaker interlocutors in Mackey and Philp (1998).

Research Question 3: Does Type of Outcome Measure Influence Observed Instructional Effectiveness?

A variety of particular outcome measures have been used within the domain to test the effect of L2 instruction. In general, because only 16% of the studies reported any form of reliability estimates for the use of outcome measures, it is not possible to assess accurately the extent to which measurement error has contributed to overall error of observation and interpretation within the domain. Furthermore, a lack of standardization among

dependent variables obscures comparisons of treatment effectiveness from study to study. Such a lack of standardization was seen not only in various response types (discussed below), but also in differences in length of measures, timing of measures, comparability of parallel test forms, variable item difficulties, construct underrepresentation and construct-irrelevant variance, and a host of other factors impacting on the construct validity of interpretations based on the use of such measures (see Messick, 1989). In response to RQ 3, therefore, there can be little doubt that the particular test or measure utilized within a given study plays a central role in observations and eventual interpretations about the effectiveness of L2 instructional treatments.

General patterns were noted across studies according to the types of responses required from learners on outcome measures, as well as in the average effect sizes associated with these different response types. On average, approximately 90% of study outcome measures required learners to utilize the L2 in accomplishing very discrete and focused linguistic tasks (meta-linguistic judgments, selected responses, constrained constructed responses), while only 10% required extended communicative use of the L2 (free constructed responses). Overall, then, observed instructional effectiveness within primary research to date has been based much more extensively on the application of explicit declarative knowledge under controlled conditions, without much requirement for fluent, spontaneous use of contextualized language.

Both selected-response (e.g., multiple choice questions on verbal conjugation) and constrained constructed-response (e.g., suppliance of a correctly conjugated verb to complete the sentence) measures were noted to have average effect sizes between 0.38 and 0.91 standard deviation units higher than meta-linguistic judgments and free constructed response measures (see Table 9). Although 95% confidence intervals overlapped for all four categories, such overlap was very limited between these two groupings. Thus, it is likely that effect sizes observed within any given study may be directly associated with the type of response required from learners on outcome measures, and associated interpretations of

study findings should be tempered by the realization that a different test type would likely have produced different results.

Although categories of outcome measures were associated with different average effect sizes, these categories were not necessarily associated with particular instructional treatment categories (see Figure 6). Primary research utilized relatively equivalent proportions of outcome measure types across each of the instructional treatment categories. Indeed, even at the individual study level, studies on average utilized at least two of the outcome measure types to interpret effectiveness of L2 instruction. Thus, although particular outcome measure types may result in very different observations about the effectiveness of a treatment, outcome measure types probably did not account for overall differences observed among different instructional treatment types in the current meta-analysis.

Research Question 4: Does Length of Instruction Influence Observed Instructional Effectiveness?

Instructional treatments were delivered over varying amounts of time in primary research studies, although treatments generally occurred within one or a few typical school class periods. Patterns of average effect sizes were noted among treatments lasting two hours or less and treatments lasting three hours or more, with shorter-term treatments resulting in an average of three tenths of a standard deviation unit greater effects than longer-term treatments (see Table 10). However, such patterns were not found to be associated with particular categories of instructional treatments. The observed differences in shorter-term versus longer-term treatment effects were likely due to the relationship among a number of study variables that are beyond the scope of the current synthesis, such as the interaction of length and intensity of instruction with target structures, the interaction between treatment and type of outcome measure, and other moderator variables.[16] To answer RQ 4, primary research will need to

treat the length and intensity of instruction systematically as experimental variables in their own right.

Research Question 5: Does Instructional Effect Last Beyond Immediate Post-experimental Observations?

An oft-raised criticism of L2 type-of-instruction research is that effects of instruction may only be short-lived at best. A number of recent primary studies have instituted one or more delayed post-tests to assess the durability of instructional effectiveness. Once again, as with other potential moderator variables, there has not been any systematic replication of this variable within accumulated primary research to date. As such, it is beyond the scope of current cumulative research findings to make any estimates regarding the durability of effects for particular treatment types or categories versus others. However, the extent to which overall instructional effects last can be interpreted with some consistency, on the basis of average findings from primary research.

In general, although the effectiveness of focused instructional treatments did seem to decrease from immediate post-test to delayed post-test observations, this decrease was on average only on the order of one fifth of a standard deviation unit (see Table 11). We also observed an accompanying decrease in the heterogeneity of effect sizes, likely due to some loss of instructional effect as well as to some target-oriented gain in control/comparison groups. It is of further interest to note that differences in the durability of effects were observed between shorter-term versus longer-term treatments. Thus, effectiveness of treatments of three hours or more in duration typically only decreased between 0.04 and 0.13 standard deviation units, while the effectiveness of treatments of less than two hours typically decreased between 0.14 and 0.41 standard deviation units. This finding may suggest a differential durability in favor of longer-term treatments, although much more careful replication of treatment duration and

timing of delayed post-tests is needed before such an observation can be interpreted with any consistency.

For the very few studies reporting sufficient data for assessing the durability of effect over several delayed post-tests, the small decrease in effect size was observed to continue and to be accompanied by continued decrease in heterogeneity of study effect sizes (see Table 12). This finding is not very trustworthy, however, owing to the very low number of studies investigating this variable. Overall, RQ 5 can be answered in the affirmative, on the basis of the evidence provided by primary research. Instructional effectiveness does seem to last beyond immediate observed effects, although it also gradually deteriorates (or control/comparison groups gradually mature).

Research Question 6: To What Extent Has Primary Research Provided Answers to These Questions?

The current synthesis has shown that a substantial number of primary research studies have utilized experimental and quasi-experimental techniques to investigate the general issue of L2 instructional effectiveness. Research questions and findings within the domain have also been similar enough to enable comparisons among studies on the basis of a general model of L2 instructional types (Doughty & Williams, 1998b), as well as on the basis of several other broad categories of shared research variables. As such, there has been considerable accumulation of primary research data related to several overarching questions about instructional effectiveness (i.e., RQs 1–5 above). However, primary research has not focused on the systematic accumulation of findings in direct response to such questions, as evinced in study designs, data analysis, and study reporting.

Study designs. Three study design features diminish the extent to which studies have contributed evidence in answer to RQs 1–5: (a) the infrequent use of true control groups, (b) the complexity of designs, and (c) the lack of replication of variables.

In the current synthesis, only 18% of 78 unique sample studies operationalized true control conditions (see Table 3). Without control conditions, wherein participants receive no treatment of any sort on the target structures, interpretations about overall change attributable to a given instructional treatment are not warranted, since there is no way of observing how much of this change may have occurred because of other factors, such as maturation, practice on the pre-test, etc. It will be recalled that rather consistent change was noted among those control groups that were operationalized in the current set of primary studies (see Figure 4).

Some L2 type-of-instruction studies have investigated the relationships among multiple independent, dependent, and moderator variables in an effort to disentangle what particular interactions among such variables may impact on instructional effectiveness (e.g., 31% of the 49 unique sample studies contributing data to the meta-analysis operationalized three or more instructional treatments within a single design). While complex designs of this sort may provide some evidence about particular interactions under particular learning circumstances, data that can be associated with a given variable (such as an instructional type) are only with great difficulty extracted from such designs. Findings from complex studies are thus very difficult to compare with findings from other studies about the same variable. Rosenthal (1991) comments on interpretations associated with multivariate data analyses and complex research designs:

> [W]e are getting quantitative answers to questions that are often—perhaps usually—hopelessly imprecise. Only rarely is one interested in knowing for any fixed-factor analysis of variance or covariance that somewhere in the thicket of *df* there lurk one or more meaningful answers to meaningful questions that we had not the foresight to ask of our data. (p. 13)

For such reasons, the American Psychological Association (1996) has taken steps to encourage the "principle of parsimony" in primary research by recommending the use of "minimally sufficient designs

and analytic strategies" (p. 2) necessary for addressing research questions (see also Cohen, 1990).[17]

Study designs in this domain may also reflect what Light and Pillemer (1984) have referred to as "the myth of the single decisive study" (p. 159). As has been observed across experimental research in the social sciences (e.g., Cohen, 1997; Rosenthal, 1991), individual primary research studies are often reported and interpreted as if they could provide definitive answers to research questions of interest to the domain. This is of course an impossibility, as, owing to the error associated with any single set of observations, individual studies can never do more than supply one additional small piece of evidence to the overall puzzle of a research question. Warranted and statistically trustworthy answers to research questions may thus only be sought through systematic reduction of sampling error via the accumulation of findings about a given variable across a range of studies.

On the whole, although motivated by common theoretical premises and associated research problems, L2 type-of-instruction research has not directly engaged in the systematic accumulation of findings (i.e., across a variety of study contexts) about research variables. Such systematicity can only be achieved by acknowledging replication as a central undertaking of primary research in cumulative scientific endeavor (Bangert-Drowns, 1986; Polio & Gass, 1997). This is not to suggest that the main undertaking of L2 type-of-instruction research should be to replicate studies to "improve" on previous research by systematically modifying variables (a common misconception). Rather, the purpose of replication should be to provide robust enough data for a domain to make trustworthy interpretations about a given variable, such as a type of instructional treatment. Such robustness can come only from the consistent operationalization of a given variable under a variety of circumstances; what gets replicated is the variable, not the study (Rosenthal, 1979b). Thus, to provide consistent answers to the questions that the field is asking, the primary study should be seen as contributing data points to a cooperative enterprise, wherein particular research variables are held constant

(replicated) across multiple studies, and findings about such variables may therefore be accumulated. Such replication for the purpose of accumulation of knowledge has been relatively unknown within the current domain (but see an impressive opus of replication in Kubota 1994, 1995a, 1995b, 1996).

Data analysis. The extent to which primary research has contributed evidence in answer to the research questions above has also been influenced by the ways in which researchers have analyzed and interpreted quantitative study findings. Of the 77 study reports reviewed in the current synthesis, 91% reported interpretations of quantitative findings according to results of statistical significance tests (see Table 6), which were therefore the second most frequently reported type of quantitative information within the domain, following only the reporting of overall study sample sizes (in 100% of the study reports). Descriptive statistics (and especially measures of dispersion) were reported less frequently, and results were interpreted using statistical significance tests far more frequently than with any other interpretive techniques (i.e., magnitude of observed effects, the strength of observed relationships, or the consistency/error of observations). The statistical significance test was thus observed to be the analytic and interpretive tool of choice within L2 type-of-instruction research (Oakes, 1986, found similar patterns in other domains of experimental research).

It is beyond the scope of the current discussion to detail the problems that may be associated with the use of statistical significance testing (see the American Psychological Association, 1994, 1996; Carver, 1978, 1993; Cohen, 1990, 1997; Frick, 1996; Harlow, Mulaik, & Steiger, 1997; Kirk, 1996; Meehl, 1997; Oakes, 1986; Rosnow & Rosenthal, 1989; Schmidt, 1996; Shaver, 1993; Snow & Wiley, 1991; Thompson, 1992, 1996, 1998; Thompson & Snyder, 1997). However, several associated problems warrant brief attention here, as these problems were noted throughout the L2 type-of-instruction domain and as they inhibit the accumulation of useful knowledge in response to its research questions.

The most obvious problem in L2 type-of-instruction research is that statistical significance test results were frequently misinterpreted as showing: (a) the presence or absence of effects or relationships, or (b) the magnitude or importance of effects or relationships. Such interpretations fail to acknowledge the role that sample size plays in tests of statistical significance. That is, statistical significance is always dependent on both the observed effect or relationship *and* the size of the study sample. Virtually any effect or relationship can be observed to be statistically significant or not, depending on the size of the sample (see Carver, 1993; Rosenthal, 1991, 1994; Schmidt, 1992; Thompson, 1994).

A second problem is that the results of statistical significance tests were frequently reported in lieu of other types of information (such as descriptive statistics). What is more, within the reporting of statistical significance tests themselves, the outcomes of the test (i.e., significant or not) were often reported in lieu of the inferential data (i.e., exact values for p, df, F, or t, and inferential statistics tables). Several studies reported no quantitative data whatsoever beyond the observation that a finding was statistically significant or not. Primary researchers also occasionally reported data (e.g., means and standard deviations) only for those comparisons found to be statistically significant while not reporting data for comparisons that were not statistically significant. The prioritization of statistical significance test results over other forms of data has thus decreased the presentation of quantitative findings in forms accessible for accurate interpretation and accumulation.

A third problem is that analysis and reporting of data in terms of statistical significance tests may also lead researchers and readers to unwarranted interpretations and conclusions about the state of findings within the domain. By way of example in the current meta-analysis, four studies were identified with comparisons which produced identical effect size estimates (d = 0.68). In one of these studies (Mackey & Philp, 1998), the effect was determined to be marginal, although not statistically significant ($p < .08$). In the second study (Cadierno, 1995), no statistically significant difference was found for the same effect

(p = .6614). Yet in the third and fourth studies (Kubota, 1994, 1996), exactly the same effect observed in the first two studies was found statistically significant ($p < .05$). Readers and reviewers attempting to interpret the findings from these studies based only on the results of statistical significance tests (i.e., two statistically significant findings and two statistically nonsignificant findings) would be led to a very different conclusion than that suggested by the actual patterns observed within each study (i.e., exactly the same magnitude of effect).

The most fundamental problem with the use of statistical significance tests in L2 type-of-instruction research is that such tests are not designed to provide answers to the primary research questions of the domain. Thus, the test of statistical significance on its own does not provide any indication of: (a) whether or not an effect or relationship was observed in the data; (b) how big or important any observed effect or relationship may have been; (c) how trustworthy or consistent any observed effect or relationship may have been; or (d) the probability that an observed effect or relationship was due to chance (see Cohen, 1990). Instead, all that the test of statistical significance indicates is whether or not an observed effect or relationship was probabilistically rare or unlikely, under the assumption that the groups being compared were randomly sampled from a single population whose parameters can be estimated on the basis of the characteristics of these groups (Carver, 1978, 1993; Cohen, 1990, 1997; Kromrey & Foster-Johnson, 1996; Schmidt & Hunter, 1997; Thompson, 1994).[18]

Fundamentally, then, a finding of statistical significance does not shed light on *why* an observed effect or relationship is rare. Thus, a statistically significant observation may result from a large observed effect or relationship, or it may simply result because sampling error was reduced enough (e.g., by sampling large numbers of subjects) so that the zero value of no difference (the null hypothesis) was not included within the given probability level (see also Frick, 1996; Kirk, 1996; Kromrey & Foster-Johnson, 1996). Obviously, not finding a statistically significant difference

does not mean that there was no effect for a given treatment (although this is exactly how it is often interpreted).

What the statistical significance test cannot reveal on its own, then, is exactly the kind of information L2 type-of-instruction research asks of its data: (a) how effective a treatment was, (b) the degree to which one treatment was more effective than another, or (c) how trustworthy interpretations may have been about a treatment or other variables. To answer such questions, other kinds of analyses must be incorporated, considering both the magnitude of the observed effect or relationship (e.g., d, η^2) and the influence of sampling error (e.g., standard errors, confidence intervals). Indeed, the calculation of effect sizes and confidence intervals provides the same probability information as that found in a statistical significance test, and it provides additional information about the size of an observed difference/relationship and the consistency of these observations (Cohen, 1990; Rosnow & Rosenthal, 1989; Rosenthal, 1994). As Meehl (1997) has pointed out, data for these analyses are always available in primary experimental research, and they are often calculated automatically along with most statistical significance tests; they are simply seldom reported or used in the interpretation of study findings.

Study reporting. In L2 type-of-instruction research, focal study variables, including independent variables, dependent variables, and moderator variables, have been infrequently reported with sufficient clarity to enable comparison with other investigations of the same variables (a quality that is essential for the accumulation of knowledge about a variable) or to enable replication of the variable in future research (see also Whittington, 1998). In the current synthesis, it was also observed that primary studies very infrequently reported in sufficient detail what actually occurred within an investigation. Thus, although researchers may describe the intended operationalization of variables, it often remains unclear, for example, whether or not an instructional treatment was actually delivered by teachers according to plan, whether or not learners reacted in intended ways, and whether or

not outcome measures elicited enough and appropriate language use to warrant interpretations.

Widespread inconsistency was also noted in the reporting of quantitative data and analyses, posing perhaps the most fundamental threat to accumulation and comparison of study findings. Basic descriptive statistics were lacking from many study reports, including group sample sizes, measures of central tendency, and, especially, measures of dispersion (e.g., standard deviations), and the reporting of inferential statistics varied among study reports (see above). Virtually none of the L2 type-of-instruction studies reported any measures of the error or consistency of their observations or the magnitude of observed effects, despite the fact that studies made direct interpretations about the effectiveness of instructional treatments.

As Polio and Gass (1997) among others have pointed out, reporting inconsistencies may be due in part to editorial policies, limited space afforded to journal articles, and the lack of clear guidelines for what should be reported. Nevertheless, although not consistently maintained within editorial practices, clear guidelines do at least exist for the reporting of quantitative data and analyses. Relative to three primary reporting problems in the current domain, the American Psychological Association (1994) is clear on what minimally should be included in a study report. In terms of inferential and descriptive statistics, it recommends:

> When reporting inferential statistics (e.g., *t* tests, *F* tests, and chi-square), include information about the obtained magnitude or value of the test, the degrees of freedom, the probability level, and the direction of the effect. Be sure to include descriptive statistics (e.g., means or medians); where means are reported, always include an associated measure of variability, such as standard deviations, variances, or mean square errors. (pp. 15–16)

In addition, it addresses problems with reporting of probability levels and the associated interpretation of observed effects:

> Neither of the two types of probability values reflects the importance (magnitude) of an effect or the strength of a

> relationship because both probability values depend on sample size. You can estimate the magnitude of the effect or the strength of the relationship with a number of measures that do not depend on sample size [. . .]. You are encouraged to provide effect-size information. (p. 18)

More recently, the American Psychological Association (1996) Task Force on Statistical Inference has concluded that "both direction and size of effect [. . .] and their confidence intervals should be provided routinely as part of the presentation [of results]" (p. 2).

Finally, the accumulation of accurate findings about the variables and research questions of interest to the domain is likely hampered by a serious bias, both among primary researchers and among editorial boards, that prioritizes the reporting of investigations that have made statistically significant observations.

Recommendations for improving research practice. In light of the various delimiting factors found in relation to research design, analysis, and reporting for investigations of L2 instructional effectiveness, we would like to offer a few specific recommendations that seem essential in order for the domain to become better able to answer its research questions:

1. Utilize simple designs that investigate only a few variables at most; interactions of variables should be investigated systematically across multiple experiments, not within single experiments.

2. Incorporate pre-tests and post-tests as well as true control groups in experimental and quasi-experimental study designs, to identify better the amount of observed effects attributable to instructional treatments.

3. Design studies with the replication of variables (not other studies) in mind; avoid the myth of the single decisive study by engaging in long-term research agendas in which a series of studies systematically provides data points about specific variables.

4. Consider the validity of dependent variables in terms of the kinds of interpretations to be based on them; estimate and report the consistency or reliability of the use of outcome measures.

5. Choose the analytic and interpretive techniques that will provide accurate answers to the research questions that are being asked; where used, interpret results of statistical significance tests appropriately.

6. For questions about the presence of an effect, the size of an effect, or the importance of an effect, calculate effect sizes (statistical significance tests will not provide answers to any of these questions).

7. Incorporate estimates of error (e.g., standard error, confidence intervals) into all quantitative analyses of experimental data.

8. Report enough data about independent, dependent, and moderator variables such that related findings may be compared with other investigations of the same variables and such that future researchers will be able to replicate these variables; include observations about what actually occurred when variables were operationalized in investigations.

9. Always report the data necessary to enable further interpretation and accumulation of study findings, including, but not limited to: means, standard deviations, and group sample sizes on all pre- and post-experimental measures (regardless of statistical significance); where used, report complete results of statistical significance tests (e.g., not just the probability levels or F or t values for findings that were observed to be statistically significant).

It is our hope that editorial practice will also give consideration to this range of issues, especially to inconsistencies in the reporting of primary research data, analyses, and interpretations, as well as to the issue of publication bias. A variety of strategies may help reduce the impact of publication bias, including: (a) requiring the

reporting of effect size estimates anywhere statistical significance tests are reported; (b) demanding correct interpretation of statistical significance tests in study reports (e.g., not allowing the reporting of "trends" towards statistical significance or "highly" significant differences); and (c) establishing a system of two-part reviews, in which reviewers are initially blind to Results sections (see Frick, 1996; Kupfersmid, 1988; Shaver, 1993; Thompson, 1994).

Conclusion

In the current study, we engaged in secondary research of a broad range of findings from primary investigations with two purposes in mind. First, we wanted to synthesize the state of experimental and quasi-experimental research methods and reporting practices within the domain of studies investigating L2 instructional effectiveness. As discussed in the previous section, we found such methods and practices to be widely variable and generally not conducive to the systematic accumulation of knowledge about particular variables. We hope that our recommendations for improving practice along these lines will be taken to heart by those conducting primary research, those reviewing and interpreting study findings, and those publishing study reports.

Our second purpose in the current study was to provide a quantitative summary of findings about several variables of general interest to L2 type-of-instruction research. In addressing this purpose, we utilized meta-analytic techniques to compare quantitative study findings on the basis of a common scale (the effect size). It was our hope that this meta-analysis would provide a precise depiction of what research thus far has found regarding L2 instructional effectiveness and several related variables. However, it should be noted that a substantial part (37%) of the body of primary research investigated did not report findings in a manner accessible for further cumulative analysis. The summary of findings discussed below must therefore be interpreted with this caveat in mind. Nevertheless, the results of the meta-analysis should offer a useful empirical context within which future

single-study findings from L2 type-of-instruction research can be more meaningfully interpreted. We turn, then, to a summary of what the available research data have to say about L2 instructional effectiveness.

In general, focused L2 instruction results in large gains over the course of an intervention. Specifically, L2 instruction of particular language forms induces substantial target-oriented change, whether estimated as pre-to-post change within experimental groups (d = 1.66) or as differences in performance between treatment and control groups on post-test measures (d = 0.96), even when the control group is exposed to and interacts with experimental materials in which the L2 form is embedded (d = 0.75). All of these average effects have been observed to differ consistently from the null hypothesis; that is, they are probabilistically rare and may therefore lead to relatively trustworthy interpretations.

The effects of L2 instruction seem durable. This can be concluded from the cumulative empirical observation that, although such effects tend to marginally decrease over time (probably as a result of learning and maturation that bring control and instructed groups closer together), it is the case that average effect sizes for delayed post-tests remain relatively large, indicating sustained differences in favor of instructed groups. However, owing to the small number of studies that have included delayed post-tests, this finding should not be interpreted as definitive.

On average, instruction that incorporates explicit (including deductive and inductive) techniques leads to more substantial effects than implicit instruction (with average effect sizes differing by 0.59 standard deviation units), and this is a probabilistically trustworthy difference. In addition, instruction that incorporates a focus on form integrated in meaning is as effective as instruction that involves a focus on forms. Thus, although both FonF and FonFS instructional approaches result in large and probabilistically trustworthy gains over the course of an investigation, the magnitude of these gains differs very little between the two instructional categories. Finally, the observed order of effectiveness

for more specific instructional types (explicit FonF > explicit FonFS > implicit FonF > implicit FonFS) is suggestive of needed future research.

Interpretation of these cumulative findings for explicit/implicit and FonF/FonFS instructional treatments should be tempered by several methodological observations. First, testing of learning outcomes usually favors explicit treatments by asking learners to engage in explicit memory tasks and/or in discrete, decontextualized L2 use. In addition, explicit treatments are typically more intense and varied than implicit treatments, and implicit treatments may require longer post-intervention observation periods for nonlinear learning curves to be detected (see also Mellow, Reeder, & Forster, 1996). Second, the essential features that supposedly distinguish FonF and FonFS instructional approaches have been inconsistently operationalized, and the wide range of actual observed effect sizes within each category suggests that the particular subtypes of instruction need to be further investigated in their own right by means of careful replication. Third, research settings vary widely, especially according to instructional contexts, number and characteristics of learner participants, and amount and intensity of instruction, all factors potentially contributing to heterogeneity in observed instructional effectiveness. These caveats notwithstanding, the current state of empirical findings indicates that explicit instruction is more effective than implicit instruction and that a focus on form and a focus on forms are equally effective.

Current cumulative knowledge also suggests that the outcome measures selected for assessing the impact of instructional treatments do lead to substantially different observations of instructional effectiveness. Namely, effects are likely to be greater in studies that employ selected-response or constrained constructed-response test formats, whereas instruction is likely to result in smaller observed effects if researchers choose to employ metalinguistic judgment response or free-response test formats. In addition, shorter instructional interventions may yield greater observed effects than do longer interventions, and the causes for

this pattern merit future research that systematically explores the relative effects of intensity and duration of treatments. Finally, up to 18% of change observed over the course of an investigation may be due to factors such as maturational effects or test practice effects. Therefore, individual researchers need to assess change in control groups and account for such change in their discussion of instructional effects and in their inferences about the actual effectiveness of particular techniques.

A more complex agenda has begun to unfold within L2 type-of-instruction research that investigates not only the relative effectiveness of particular instructional techniques but also the potential impact of a range of moderator variables (e.g., learner factors such as aptitude, age, and learning style; linguistic factors, such as the relative structural complexity of L2 forms; cognitive factors, such as learner developmental readiness, degree of noticing; and pedagogical factors, such as timing, duration, and intensity of instruction, and integration of interventions within the language curriculum). Furthermore, research is being carried out with widely differing populations (e.g., university versus elementary students) and in widely varying instructional contexts (e.g., classrooms, laboratories). For this new research agenda to be adequately investigated, the L2 type-of-instruction research domain will need to agree upon consistent empirical operationalizations of its central constructs in the form of variables that may be replicated across such populations and contexts. In addition, researchers will need to turn to more rigorous practices for experimental and quasi-experimental designs, and they will need to engage in careful, long-term examination of the central questions and constructs that motivate research into the effectiveness of L2 instruction.

Revised version accepted 26 November 1999

Notes

[1]Proponents of the so-called non-interface position hold that true linguistic competence remains unaffected by rule presentation and negative feedback (see Krashen, 1985, 1999; Schwartz, 1993; Paradis, 1994; Young-Scholten, 1999).

[2]Rosenthal (1991, p. 61) presents the following example of how primary research findings may be misinterpreted by looking only at the results of statistical significance tests:

> For example, Smith may report a significant effect of some social intervention only to have Jones publish a rebuttal demonstrating that Smith was wrong in her claim. A closer look at both their results may show the following:
>
> Smith's study: $t(78) = 2.21, p < .05, d = .50, r = .24$
> Jones's study: $t(18) = 1.06, p > .30, d = .50, r = .24$

The actual relationship observed in each of the two studies is exactly the same (i.e., d, the observed effect size within each study, and r, the observed correlation between variables within each study, are exactly the same for the two studies). In fact, the only difference between the two studies, and the source for the differing t test results, is the fact that Smith's study has a larger sample size ($n = 79$) than Jones's study ($n = 19$). In a vote-counting review, the reviewer would conclude, as Jones did, that findings about the particular social intervention thus far are inconsistent, since one study observed statistically significant results in favor of the intervention and a second study observed no statistically significant results. The fact is, however, that both studies offer support for the hypothetical social intervention, as both observed exactly the same positive effect ($d = .50$) in favor of the intervention, regardless of the fact that one study observed this effect on a larger sample.

[3]Journals reviewed: *Applied Linguistics, Applied Psycholinguistics, Applied Language Learning, Canadian Modern Language Review, Foreign Language Annals, JALT Journal, Language Learning, Language Teaching Research, Modern Language Journal, RELC Journal, Second Language Research, Studies in Second Language Acquisition, System,* and *TESOL Quarterly*.

[4]Although the literature search resulted in a high frequency of redundant identifications, it was hoped that such redundancy would enable the exhaustive identification of all relevant study reports. Of course, no matter how exhaustive these search techniques, it is likely that other study reports warranting inclusion were not identified. An exact list of all study reports included in the current synthesis is thus provided in the References section with the hope that readers and future reviewers will be able to identify any such missing reports. Unfortunately, space constraints preclude listing the 250-plus study reports identified in the literature search as potentially relevant.

[5]Our thanks to Peter Robinson for tracking down the final two of many retrieved study reports.

[6]Reporting quality is likely to improve over time in any research domain (see, e.g., Orwin, 1983).

[7]DeKeyser (1998) has pointed out, "It is rather uncontroversial that pronunciation is relatively immune to all but the most intensive formS-focused

treatments, whereas large amounts of vocabulary can be acquired with very little focus on form" (p. 43).

[8]Our thanks go to Cathy Doughty and to students in her graduate seminar at the University of Hawaii, fall semester, 1997, for their feedback during early stages of the research synthesis.

[9]Obviously, the two senses of explicitness should be thought of as forming a continuum, rather than a dichotomy, with explicit treatments ranging from the more deductive to the more inductive.

[10]Indeed, DeKeyser and Sokalski (1996, pp. 625–626) explain that both comprehension and production treatments contained the same blend of meaningful and mechanical exercises, thus counterbalancing the requirements for a focus on meaning and forms in both treatments. This strategy, which is a sound experimental practice to make treatments comparable and reduce potential confounds, also makes obvious that the integration of form and meaning was not considered a psycholinguistic requirement for effective instruction. Thus, both treatments can be viewed as more FonFS than FonF.

[11]It should be emphasized here that studies were coded according to the explicit/implicit and FonF/FonFS/FonM distinctions entirely on the basis of the evidence provided within study reports (see Appendix A). We have no doubt that there will be those who disagree with our categorizations of certain instructional treatments within the current synthesis. It is our hope that any such disagreements will lead primary researchers as well as those interested in the research domain to consider the importance of definitional criteria for and the concomitant operationalization of independent variables, as well as the adequate reporting of such variables. Should readers disagree with particular codings found here, it is our hope that such disagreement will give rise to careful consideration of the source of this disagreement, with particular emphasis on two issues. First, as coding decisions were based on study reports, how (and how well) were instructional treatments reported in these sources? Was the reporting sufficient to convey the intended operationalization of the independent variable? Second, did the actual instructional treatments involve an appropriate operationalization of the intended independent variable? Were intended instructional treatments confounded by the addition of other instructional features? We hope that answers to such questions will lead to increased precision in both the operationalization of instructional treatments and the reporting of independent variables.

[12]As the meta-analysis literature has pointed out (Rosenthal, 1994), Equation 3 should not be used with F values that are based on omnibus F tests with $df > 1$ in the numerator. The resulting effect size estimate is not the same as it would be when using F or t from a direct contrast. In the current synthesis, a number of studies were found to report F-test comparisons without sufficient information to determine whether or not the resulting F value was based on a direct contrast. There are methods available for correctly calculating between-groups effects from multivariate designs with embedded comparisons (e.g., Glass et al., 1981; Rosenthal & Rosnow, 1991; Rosenthal &

Rubin, 1986; Seifert, 1991). Such calculations require a series of data transformations based on the reporting of extensive information, typically in the form of inferential statistics tables (e.g., ANOVA tables). However, as Cooper (1998) has pointed out, "primary researchers rarely report their results in enough detail to carry out the needed transformations" (p. 93). As such tables were almost never adequately reported in the current research domain, calculation of effect sizes by using the results of multivariate analyses was not undertaken.

[13]Several studies reported descriptive or inferential statistics that could not be precisely linked to individual study groups; these studies were not included in the quantitative meta-analysis.

[14]This formula for the 95% confidence interval can be approximated with Equation 5 for sample sizes greater than 30 (e.g., Rosenthal, 1995, p. 187), where the *t* distribution approaches a constant value:

$$CI = d \pm 2\frac{SD}{\sqrt{k}} \tag{5}$$

[15]If the learner failed to self-correct after the provision of an initial clarification request, a brief grammatical explanation would be offered by the teacher-researcher, followed by a new invitation to self-correct, which would be followed up by a full grammatical explanation if the learner failed to self-correct after the second opportunity.

[16]As one anonymous reviewer pointed out, the inevitable loss of control over experimental conditions in longer-duration treatments must also contribute to smaller observed effect sizes.

[17]A number of readers of this manuscript have suggested that studies employing simple research designs by focusing only on several particular variables may not capture the complexity of the educational contexts within which L2 instructional interventions typically occur. While we agree that language education is a complex undertaking, and that a range of learner, context, treatment, and other variables related to an investigation should be carefully observed and rigorously reported, we do not agree that the single experimental study that occurs within a typical instructional context can provide sufficient data for testing or accurately interpreting interactions among such a range of variables. We suggest that interaction effects are most accurately interpreted at the meta-analytic level, and that interactions can best be investigated via the accumulation of findings across a range of studies. Of course, such interpretation may be undertaken with accuracy only after particular variables have been carefully observed (and carefully reported) across multiple research and instructional contexts.

[18]Carver (1978) has summarized:

> Statistical significance simply means statistical rareness. Results are "significant" from a statistical point of view because they occur very rarely in random sampling under the conditions of the

> null hypothesis. A statistically significant mean difference between two research groups at the .05 level indicates the following: if we assume that the two research groups are random samples representing the same hypothetical population which has properties that can be estimated from properties of the groups themselves, and if we assume that we sampled 100 sets of two groups from this same hypothetical population, then we would expect to find the mean difference between the two research groups to be larger than 95 of the 100 sampled from the hypothetical population. A statistically significant result means that the probability is low that we would get the type of result we got, given that the null hypothesis is true. (p. 383)

References

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

American Psychological Association. (1994). *Publication manual of the American Psychological Association* (4th ed.). Washington, DC: Author.

American Psychological Association. (1996). *Task force on statistical inference initial report* [html document]. Available: [http://www.apa.org/science/tfsi.html]

Atkinson, R., Furlong, M., & Wampold, B. (1982). Statistical significance, reviewer evaluations, and the scientific process: Is there a (statistically) significant relationship? *Journal of Counseling Psychology*, *29*, 189–194.

Bangert-Drowns, R. L. (1986). Review of developments in meta-analytic method. *Psychological Bulletin*, *99*, 388–399.

Begg, C. (1994). Publication bias. In H. Cooper & L. Hedges (Eds.), *Handbook of research synthesis* (pp. 399–409). New York: Russell Sage Foundation.

*Billmyer, K. A. (1990). "I really like your lifestyle": ESL learners learning how to compliment. *Penn Working Papers in Educational Linguistics*, *6*(2), 31–48. (ERIC ED 335937)

---

*Asterisks mark the 77 study reports that constituted the body of research synthesized in the current study..

**Double asterisks mark the 45 reports showing sufficient interpretable data to calculate an effect size estimate for in the quantitative meta-analysis.

**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)

Bushman, B. (1994). Vote-counting procedures in meta-analysis. In H. Cooper & L. Hedges (Eds.), *The handbook of research synthesis* (pp. 193–214). New York: Russell Sage Foundation.

Cadierno, T. (1992). *Explicit instruction in grammar: A comparison of input-based and output-based instruction in second language acquisition*. Unpublished doctoral dissertation, University of Illinois.

**Cadierno, T. (1995). Formal instruction from a processing perspective: An investigation into the Spanish past tense. *The Modern Language Journal*, *79*, 179–193.

Cameron, J., & Pierce, W. (1994). The debate about rewards and intrinsic motivation: Protests and accusations do not alter the results. *Review of Educational Research*, *66*, 39–51.

**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.

**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.

Carver, R. (1978). The case against statistical significance testing. *Harvard Educational Review*, *48*, 389–399.

Carver, R. (1993). The case against statistical significance testing, revisited. *Journal of Experimental Education*, *61*, 287–292.

Chaudron, C. (1977). A descriptive model of discourse in the corrective treatment of learners' errors. *Language Learning*, *27*, 29–46.

Chaudron, C. (1985). Intake: On models and methods for discovering learners' processing of input. *Studies in Second Language Acquisition*, *7*, 1–14.

Chaudron, C. (1988). *Second language classrooms: Research on teaching and learning*. Cambridge: Cambridge University Press.

Chaudron, C. (1998, April). Contrasting approaches to classroom research: Qualitative and quantitative analysis of language use and learning. Paper presented at the XVI Congreso Nacional de la Asociación Española de Lingüística Aplicada, Logroño, Spain.

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.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, *45*, 1304–1312.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155–159.

Cohen, J. (1994). The earth is round (r < .05). *American Psychologist*, *49*, 997–1003.

Cohen, J. (1997). The earth is round (r < .05). In L. Harlow, S. Mulaik, & J. Steiger (Eds.), *What if there were not significance tests*? (pp. 21–36). Mahwah, NJ: Lawrence Erlbaum Associates.

Cook, T. (1992). *Meta-analysis for explanation: A case book*. New York: Russell Sage Foundation.

Cook, T., & Campbell, D. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Boston: Houghton Mifflin Co.

Cooper, H. (1989). *Homework*. New York: Longman.

Cooper, H. (1998). *Synthesizing research: A guide for literature reviews* (3rd ed.). Thousand Oaks, CA: Sage.

Cooper, H., DeNeve, K., & Charlton, K. (1997). Finding the missing science: The fate of studies submitted for review by a human subjects committee. *Psychological Methods*, *2*, 447–452.

Cooper, H., & Hedges, L. V. (Eds.). (1994a). *The handbook of research synthesis*. New York: Russell Sage Foundation.

Cooper, H., & Hedges, L. V. (1994b). Research synthesis as a scientific enterprise. In H. Cooper & L. Hedges (Eds.), *The handbook of research synthesis* (pp. 3–14). New York: Russell Sage Foundation.

**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. (1994). How implicit can adult second language learning be? *AILA Review*, *11*, 83–96.

**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: Automatizing second language morphosyntax. *Studies in Second Language Acquisition*, *19*, 195–221.

DeKeyser, R. M. (1998). Beyond focus on form: Cognitive perspectives on learning and practicing second language grammar. In C. Doughty & J. Williams (Eds.), *Focus on form in classroom second language acquisition* (pp. 42–63). Cambridge: Cambridge University Press.

**DeKeyser, R. M., & Sokalski, K. J. (1996). The differential role of comprehension and production practice. *Language Learning*, *46*, 613–642. [Two studies]

**Doughty, C. (1991). Second language instruction does make a difference: Evidence from an empirical study of ESL relativization. *Studies in Second Language Acquisition*, *13*, 431–469.

Doughty, C. (1997, October). *Meeting the criteria of focus on form*. Paper presented at the 17th Second Language Research Forum, Michigan State University, East Lansing, MI.

Doughty, C. (1998). Acquiring competence in a second language: Form and function. In H. Byrnes (Ed.), *Learning foreign and second languages* (pp. 128–156). New York: Modern Language Association of America.

*Doughty, C., & Varela, E. (1998). Communicative focus on form. In C. Doughty & J. Williams (Eds.), *Focus on form in classroom second language acquisition* (pp. 114–138). Cambridge: Cambridge University Press.

Doughty, C., & Williams, J. (Eds.). (1998a). *Focus on form in classroom second language acquisition*. Cambridge: Cambridge University Press.

Doughty, C., & Williams, J. (1998b). Issues and terminology. In C. Doughty & J. Williams (Eds.), *Focus on form in classroom second language acquisition* (pp. 1–12). Cambridge: Cambridge University Press.

Doughty, C., & Williams, J. (1998c). Pedagogical choices in focus on form. In C. Doughty & J. Williams (Eds.), *Focus on form in classroom second language acquisition* (pp. 197–261). Cambridge: Cambridge University Press.

Dubin, R., & Taveggia, T. (1968). *The teaching–learning paradox: A comparative analysis of college teaching methods*. Eugene, OR: University of Oregon Press.

*Eckman, F., Bell, L., & Nelson, D. (1988). On the generalization of relative clause instruction in the acquisition of English as a second language. *Applied Linguistics*, *9*, 1–20.

*Ellis, N. (1993). Rules and instances in foreign language learning: Interactions of implicit and explicit knowledge. *European Journal of Cognitive Psychology*, *5*, 289–319.

Ellis, N. (1994). Introduction: Implicit and explicit language learning—An overview. In N. Ellis (Ed.), *Implicit and explicit learning of languages* (pp. 1–31). London: Academic Press.

Ellis, N. C., & Laporte, N. (1997). Contexts of acquisition: Effects of formal instruction and naturalistic exposure on second language acquisition. In A. de Groot & J. Kroll (Eds.), *Tutorials in bilingualism: Psycholinguistic perspectives* (pp. 53–83). Hillsdale, NJ: Lawrence Erlbaum.

*Ellis, R. (1984). Can syntax be taught? A study of the effects of formal instruction on the acquisition of WH questions by children. *Applied Linguistics*, *5*, 138–155.

Ellis, R. (1994). *The study of second language acquisition*. Oxford: Oxford University Press.

Ellis, R. (1998). Teaching and research: Options in grammar teaching. *TESOL Quarterly*, *32*, 39–60.

**Ellis, R., Rosszell, H., & Takashima, H. (1994). Down the garden path: Another look at negative feedback. *JALT Journal*, *16*, 9–24.

*Fotos, S. (1993). Consciousness raising and noticing through focus on form: Grammar task performance versus formal instruction. *Applied Linguistics*, *14*, 385–407. [Same study as Fotos, 1994]

*Fotos, S. (1994). Integrating grammar instruction and communicative language use through grammar consciousness-raising tasks. *TESOL Quarterly*, *28*, 323–351. [Same study as Fotos, 1993]

**Fotos, S., & Ellis, R. (1991). Communicating about grammar: A task-based approach. *TESOL Quarterly*, *25*, 605–628. [Two studies]

Frick, R. (1996). The appropriate use of null hypothesis testing. *Psychological Methods*, *1*, 379–390.

*Gass, S. (1982). From theory to practice. In M. Hines & W. Rutherford (Eds.), *On TESOL '81* (pp. 129–139). Washington, DC: Teachers of English to Speakers of Other Languages.

Gass, S. M., & Varonis, E. M. (1994). Input, interaction, and second language production. *Studies in Second Language Acquisition*, *16*(3), 283–302.

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

Glass, G., McGaw, B., & Smith, M. (1981). *Meta-analysis in social research*. Beverly Hills, CA: Sage.

Greenhouse, J., & Iyengar, S. (1994). Sensitivity analysis and diagnostics. In H. Cooper & L. Hedges (Eds.), *Handbook of research synthesis* (pp. 383–398). New York: Russell Sage Foundation.

Greenwald, A. (1975). Consequences of prejudices against the null hypothesis. *Psychological Bulletin*, *82*, 1–20.

*Hamilton, R. (1994). Is implicational generalization unidirectional and maximal? Evidence from relativization instruction in a second language. *Language Learning*, *44*, 123–157.

Harley, B. (1988). Effects of instruction on SLA: Issues and evidence. *Annual Review of Applied Linguistics*, *9*, 165–178.

**Harley, B. (1989). Functional grammar in French immersion: A classroom experiment. *Applied Linguistics*, *10*, 331–359.

Harley, B. (1994). Appealing to consciousness in the L2 classroom. *AILA Review*, *11*, 57–68.

*Harley, B. (1998). The role of focus on form in promoting child L2 acquisition. In C. Doughty & J. Williams (Eds.), *Focus on form in classroom second language acquisition* (pp. 156–174). Cambridge: Cambridge University Press.

Harlow, L., Mulaik, S., & Steiger, J. (Eds.). (1997). *What if there were not significance tests?* Mahwah, NJ: Lawrence Erlbaum.

Hedges, L., & Olkin, I. (1980). Vote-counting methods in research synthesis. *Psychological Bulletin*, *88*, 359–369.

Hedges, L., & Olkin, I. (1985). *Statistical methods for meta-analysis*. New York: Academic Press.

*Herron, C. (1991). The garden path correction strategy in the foreign language classroom. *The French Review*, *64*, 966–977.

**Herron, C., & Tomasello, M. (1988). Learning grammatical structures in foreign language: Modelling versus feedback. *The French Review*, *61*, 910–922.

*Herron, C., & Tomasello, M. (1992). Acquiring grammatical structures by guided instruction. *The French Review*, *65*, 708–718.

*House, J. (1996). Developing pragmatic fluency in English as a foreign language: Routines and metapragmatic awareness. *Studies in Second Language Acquisition*, *18*, 225–252.

**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]

Hulstijn, J. H. (1997). Second language acquisition research in the laboratory: Possibilities and limitations. *Studies in Second Language Acquisition*, *19*, 131–144.

Hulstijn, J. H., & de Graaff, R. (1994). Under what conditions does explicit knowledge of a second language facilitate the acquisition of implicit knowledge? A research proposal. *AILA Review*, *11*, 97–112.

Hunter, J., & Schmidt, F. (1990). *Methods of meta-analysis*. Newbury Park, CA: Sage.

**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.

Kasper, G. (1998). Interlanguage pragmatics. In H. Byrnes (Ed.), Learning foreign and second languages (pp. 183–208). New York: Modern Language Association of America.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement*, *56*, 746–759.

Krashen, S. (1985). *The input hypothesis*. London: Longman.

Krashen, S. (1999). Seeking a role for grammar: A review of some recent studies. *Foreign Language Annals*, *32*, 245–257.

Kromrey, J., & Foster-Johnson, L. (1996). Determining the efficacy of intervention: The use of effect sizes for data analysis in single-subject research. *The Journal of Experimental Education*, *65*, 73–93.

**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)

Kupfersmid, J. (1988). Improving what is published: A model in search of an editor. *American Psychologist*, *43*, 635–642.

*Leeman, J., Arteagoitia, I., Fridman, B., & Doughty, C. (1995). Integrating attention to form with meaning: Focus on form in content-based Spanish instruction. In R. Schmidt (Ed.), *Attention and awareness in foreign language learning* (Technical Report No. 9) (pp. 217–258). Honolulu, HI: University of Hawai'i, Second Language Teaching & Curriculum Center.

**Leow, R. P. (1997). Attention, awareness, and foreign language behavior. *Language Learning*, *47*, 467–506.

**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.

Light, R., & Pillemer, D. (1984). *Summing up: The science of reviewing research*. Cambridge, MA: Harvard University Press.

Light, R., & Smith, P. (1971). Accumulating evidence: procedures for resolving contradictions among different studies. *Harvard Educational Review*, *41*, 429–471.

Lightbown, P. (1985). Can language acquisition be altered by instruction? In K. Hyltenstam & M. Pienemann (Eds.), *Modelling and assessing second language acquisition* (pp. 101–112). Clevedon, Avon: Multilingual Matters.

Lightbown, P. (1998). The importance of timing in focus on form. In C. Doughty & J. Williams (Eds.), *Focus on form in classroom second language acquisition* (pp. 177–196). Cambridge: Cambridge University Press.

Lightbown, P., & Spada, N. (1990). Focus on form and corrective feedback in communicative language teaching: Effects on second language learning. *Studies in Second Language Acquisition*, *12*, 429–448.

*Lightbown, P., Spada, N., & Wallace, R. (1980). Some effects of instruction on child and adolescent ESL learners. In S. Krashen & R. Scarcella (Eds.), *Research in second language acquisition* (pp. 162–171). Rowley, MA: Newbury House.

*Linnell, J. D. (1991). Instruction or interaction? A study of the acquisition of modals by beginning non-native speakers. *Penn Working Papers in Educational Linguistics*, *7*(2), 83–92. (ERIC ED 341249)

Lipsey, M. (1994). Identifying potentially interesting variables and analysis opportunities. In H. Cooper & L. Hedges (Eds.), *The handbook of research synthesis* (pp. 111–124). New York: Russell Sage Foundation.

Lipsey, M., & Wilson, D. (1993). The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist*, *48*, 1181–1209.

Long, M. H. (1983). Does second language instruction make a difference? A review of the research. *TESOL Quarterly*, *17*, 359–382.

Long, M. H. (1988). Instructed interlanguage development. In L. Beebe (Ed.), *Issues in second language acquisition: Multiple perspectives* (pp. 115–141). Rowley, MA: Newbury House.

Long, M. H. (1991a). The design and psycholinguistic motivation of research on foreign language learning. In B. F. Freed (Ed.), *Foreign language acquisition research and the classroom* (pp. 309–320). Lexington, MA: D. C. Heath.

Long, M. H. (1991b). Focus on form: A design feature in language teaching methodology. In K. de Bot, R. Ginsberg, & C. Kramsch (Eds.), *Foreign language research in cross-cultural perspective* (pp. 39–52). Amsterdam: John Benjamins.

Long, M. H. (1997, March). *Focus on form in task-based language teaching*. Presentation at the Fourth Annual McGraw-Hill Teleconference in Second Language Teaching. Available: [http://www.mhhe.com/socscience/foreignlang/top.htm]

Long, M. H., & Crookes, G. (1992). Three approaches to task-based syllabus design. *TESOL Quarterly*, *26*, 27–56.

Long, M. H., & Crookes, G. (1993). Units of analysis in syllabus design: The case for task. In G. Crookes & S. Gass (Eds.), *Tasks in a pedagogical context: Integrating theory and practice* (pp. 9–56). Clevedon, Avon: Multilingual Matters.

**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. H., & Robinson, P. (1998). Focus on form: Theory, research, and practice. In C. Doughty & J. Williams (Eds.), *Focus on form in classroom second language acquisition* (pp. 15–41). Cambridge: Cambridge University Press.

**Loschky, L. (1994). Comprehensible input and second language acquisition: What is the relationship? *Studies in Second Language Acquisition*, *16*, 303–323.

Loschky, L., & Bley-Vroman, R. (1993). Grammar and task-based methodology. In G. Crookes & S. Gass (Eds.), *Tasks and language learning* (pp. 123–167). Clevedon, Avon: Multilingual Matters.

Lyster, R. (1990). The role of analytic language teaching in French immersion programs. *The Canadian Modern Language Review*, *47*, 159–176.

**Lyster, R. (1994). The effect of functional-analytic teaching on aspects of French immersion students' sociolinguistic competence. *Applied Linguistics*, *15*, 263–287.

Lyster, R. (1998). Negotiation of form, recasts, and explicit correction in relation to error types and learner repair in immersion classrooms. *Language Learning*, *48*, 183–218.

**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.

Matt, G., & Cook, T. (1994). Threats to the validity of research synthesis. In H. Cooper & L. Hedges (Eds.), *The handbook of research synthesis* (pp. 503–520). New York: Russell Sage Foundation.

McLaughlin, B. (1990). Restructuring. *Applied Linguistics*, *11*, 1–16.

McLaughlin, B., & Heredia, R. (1996). Information-processing approaches to research on second language acquisition and use. In W. C. Ritchie & T. K. Bhatia (Eds.), *Handbook of second language acquisition* (pp. 213–228). New York: Academic Press.

Meehl, P. (1991). Why summaries of research on psychological theories are often uninterpretable. In R. Snow & D. Wiley (Eds.), *Improving inquiry in social science* (pp. 13–60). Hillsdale, NJ: Lawrence Erlbaum Associates.

Meehl, P. (1997). The problem is epistemology, not statistics: Replace significance tests by confidence intervals and quantify accuracy of risky numerical predications. In L. Harlow, S. Mulaik, & J. Steiger (Eds.), *What if there were not significance tests*? (pp. 393–425). Mahwah, NJ: Lawrence Erlbaum Associates.

Mellow, D., Reeder, K., & Forster, E. (1996). Using time-series research designs to investigate the effects of instruction on SLA. *Studies in Second Language Acquisition*, *18*, 325–350.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 13–103). New York: American Council on Education and Macmillan.

**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.

*Nagata, N., & Swisher, M. V. (1995). A study of consciousness-raising by computer: The effect of metalinguistic feedback on second language learning. *Foreign Language Annals*, *28*, 337–347. [Same study as Nagata, 1993]

*Nobuyoshi, J., & Ellis, R. (1993). Focused communication tasks and second language acquisition. *ELT Journal*, *47*, 203–210.

Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioral sciences*. New York: Wiley.

Orwin, R. G. (1983). *The influence of reporting quality in primary studies on meta-analytic outcomes: A conceptual framework and reanalysis*. Unpublished doctoral dissertation, Northwestern University.

Orwin, R. (1994). Evaluating coding decisions. In H. Cooper & L. Hedges (Eds.), *The handbook of research synthesis* (pp. 139–162). New York: Russell Sage Foundation.

Paradis, M. (1994). Neurolinguistic aspects of implicit and explicit memory: Implications for bilingualism and SLA. In N. C. Ellis (Ed.), Implicit and explicit learning of languages (pp. 393–419). London: Academic Press.

Petrosino, A. (1995). Specifying inclusion criteria for a meta-analysis. *Evaluation Review*, *19*, 274–293.

Pienemann, M. (1984). Psychological constraints on the teachability of languages. *Studies in Second Language Acquisition*, *6*, 186–214.

Pienemann, M. (1989). Is language teachable? *Applied Linguistics*, *10*, 52–79.

Pigott, T. (1994). Methods of handling missing data in research synthesis. In H. Cooper & L. Hedges (Eds.), *The handbook of research synthesis* (pp. 163–176). New York: Russell Sage Foundation.

Polio, C., & Gass, S. (1997). Replication and reporting: A commentary. *Studies in Second Language Acquisition*, *19*, 499–508.

Redfield, D. L., & Rousseau, E. W. (1981). A meta-analysis of experimental research on teacher questioning behavior. *Review of Educational Research*, *51*, 237–245.

Robb, T., Ross, S., & Shortreed, I. (1986). Salience of feedback on error and its effect on EFL writing quality. *TESOL Quarterly*, *20*, 83–95.

Robey, R. R. (1998). A meta-analysis of clinical outcomes in the treatment of aphasia. *Journal of Speech, Language, and Hearing Research*, *41*, 172–187.

Robinson, P. (1995a). Aptitude, awareness, and the fundamental similarity of implicit and explicit second language learning. In R. Schmidt (Ed.), *Attention and awareness in foreign language learning* (Technical Report No. 9) (pp. 303–358). Honolulu, HI: University of Hawai'i, Second Language Teaching & Curriculum Center.

Robinson, P. (1995b). Attention, memory, and the "noticing" hypothesis. *Language Learning*, *45*, 283–331.

**Robinson, P. (1996a). *Consciousness, rules, and instructed second language acquisition*. New York: Peter Lang. [Same study as Robinson, 1996b]

*Robinson, P. (1996b). Learning simple and complex second language rules under implicit, incidental, rule-search and instructed conditions. *Studies in Second Language Acquisition*, *18*, 27–68. [Same study as Robinson, 1996a]

**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. (1979a). The "file drawer problem" and tolerance for null results. *Psychological Bulletin*, *86*, 638–641.

Rosenthal, R. (1979b). Replications and their relative utility. *Replications in Social Psychology*, *1*, 15–23.

Rosenthal, R. (1983). Assessing the statistical and social importance of the effects of psychotherapy. *Journal of Counsulting and Clinical Psychology*, *51*, 4–13.

Rosenthal, R. (1991). *Meta-analytic procedures for social research*. Newbury Park, CA: Sage.

Rosenthal, R. (1994). Parametric measures of effect size. In H. Cooper & L. Hedges (Eds.), *Handbook of research synthesis* (pp. 231–244). New York: Russell Sage Foundation.

Rosenthal, R. (1995). Writing meta-analytic reviews. *Psychological Bulletin*, *118*, 183–192.

Rosenthal, R., & Rosnow, R. L. (1991). *Essentials of behavioral research: Mehods and data analysis* (2nd ed). New York: McGraw-Hill.

Rosenthal, R., & Rubin, D. (1982). Further meta-analytic procedures for assessing cognitive gender differences. *Journal of Educational Psychology*, *74*, 708–712.

Rosenthal, R., & Rubin, D. (1986). Meta-analytic procedures for combining studies with multiple effect sizes. *Psychological Bulletin*, *99*, 400–406.

Rosnow, R., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist*, *44*, 1276–1284.

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

Sahari, M. (1997). Elaboration as a text-processing strategy: A meta-analytic review. *RELC Journal*, *28*, 15–27.

**Salaberry, M. R. (1997). The role of input and output practice in second language acquisition. *The Canadian Modern Language Review*, *53*, 422–451.

Schmidt, F. (1992). What do data really mean?: Research findings, meta-analysis, and cumulative knowledge in psychology. *American Psychologist*, *47*, 1173–1181.

Schmidt, F. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for the training of researchers. *Psychological Methods*, *1*, 115–129.

Schmidt, F., & Hunter, J. (1997). Eight common but false objections to the discontinuation of significance testing in the analysis of research data. In L. Harlow, S. Mulaik, & J. Steiger (Eds.), *What if there were not significance tests*? (pp. 37–64). Mahwah, NJ: Lawrence Erlbaum Associates.

Schmidt, R. (1993). Awareness and second language acquisition. *Annual Review of Applied Linguistics*, *13*, 206–226.

Schmidt, R. (1997, October). *There is no learning without attention*. Paper presented at the 17th Second Language Research Forum, Michigan State University, East Lansing, MI.

**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.

Seifert, T. L. (1991). Determining effect sizes in various experimental designs. *Educational and Psychological Measurement*, *51*, 341–347.

Shadish, W., & Haddock, K. (1994). Combining estimates of effect size. In H. Cooper & L. Hedges (Eds.), *The handbook of research synthesis* (pp. 261–282). New York: Russell Sage Foundation.

*Shaffer, C. (1989). A comparison of inductive and deductive approaches to teaching foreign languages. *The Modern Language Journal*, *73*, 395–403.

Sharwood Smith, M. (1981). Consciousness-raising and second language acquisition theory. *Applied Linguistics*, *2*, 159–168.

Sharwood Smith, M. (1993). Input enhancement in instructed SLA: Theoretical based. *Studies in Second Language Acquisition*, *15*, 165–179.

Shaver, J. (1993). What statistical significance testing is, and what it is not. *Journal of Experimental Education*, *61*, 293–316.

*Shook, D. J. (1994). FL/L2 reading, grammatical information, and the input-to-intake phenomenon. *Applied Language Learning*, *5*, 57–93.

Skehan, P. (1998). *A cognitive approach to language learning*. Oxford: Oxford University Press.

Smith, M., & Glass, G. (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist*, *32*, 732–760.

Snow, R., & Wiley, D. (Eds.). (1991). *Improving inquiry in social science*. Hillsdale, NJ: Lawrence Erlbaum.

Spada, N. (1997). Form-focussed instruction and second language acquisition: A review of classroom and laboratory research. *Language Teaching*, *29*, 1–15.

*Spada, N., & Lightbown, P. M. (1993). Instruction and the development of questions in the L2 classroom. *Studies in Second Language Acquisition*, *15*, 205–221. [Same study sample as White, Spada, Lightbown, & Ranta, 1991]

Stock, W. (1994). Systematic coding for research synthesis. In H. Cooper & L. Hedges (Eds.), *The handbook of research synthesis* (pp. 125–138). New York: Russell Sage Foundation.

Swain, M. (1998). Focus on form through conscious reflection. In C. Doughty & J. Williams (Eds.), Focus on form in classroom second language acquisition (pp. 64–81). Cambridge: Cambridge University Press.

Taveggia, T. (1974). Resolving research controversy through empirical cumulation. *Sociological Methods and Research*, *2*, 395–407.

Terrell, T. (1991). The role of grammar instruction in a communicative approach. *Modern Language Journal*, *75*, 52–63.

Thomas, M. (1994). Assessment of L2 proficiency in second language acquisition research. *Language Learning*, *44*, 307–336.

Thompson, B. (1992). Two and one-half decades of leadership in measurement and evaluation. *Journal of Counseling and Development*, *70*, 434–438.

Thompson, B. (1994). Guidelines for authors. *Educational and Psychological Measurement*, *54*, 837–847.

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher*, *25*(2), 26–30.

Thompson, B. (1998, April). *Five methodological errors in educational research: The pantheon of statistical significance and other faux pas*. Invited

address at the annual meeting of the American Educational Research Association, San Diego, CA.

Thompson, B., & Snyder, P. (1997). Statistical significance testing practices. *The Journal of Experimental Education*, *66*, 75–83.

*Tomasello, M., & Herron, C. (1988). Down the garden path: Inducing and correcting overgeneralization errors in the foreign language classroom. *Applied Psycholinguistics*, *9*, 237–246.

Tomasello, M., & Herron, C. (1989). Feedback for language transfer errors: The garden path technique. *Studies in Second Language Acquisition*, *11*, 385–395.

Tomlin, R., & Villa, V. (1994). Attention in cognitive science and second language acquisition. *Studies in Second Language Acquisition*, *16*, 183–204.

*Trahey, M. (1996). Positive evidence in second language acquisition: Some long-term effects. *Second Language Research*, *12*, 111–139. [Same study as Trahey & White, 1993]

*Trahey, M., & White, L. (1993). Positive evidence and preemption in the second language classroom. *Studies in Second Language Acquisition*, *15*, 181–204. [Same study as Trahey, 1996]

Truscott, J. (1996). Review article: The case against grammar correction in L2 writing classes. *Language Learning*, *46*, 327–369.

**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. (1988). How juries get hung: Problems with the evidence for a focus on form in teaching. *Language Learning*, *38*, 243–260.

VanPatten, B. (1994). Evaluating the role of consciousness in second language acquisition: Terms, linguistic features and research methodology. *AILA Review*, *11*, 27–36.

**VanPatten, B., & Cadierno, T. (1993a). Explicit instruction and input processing. *Studies in Second Language Acquisition*, *15*, 225–241.

*VanPatten, B., & Cadierno, T. (1993b). Input processing and second language acquisition: A role for instruction. *The Modern Language Journal*, *77*, 45–57.

**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.

Wampold, B., Mondin, G., Moody, M., Stich, F., Benson, K., & Ahn, H. (1997). A meta-analysis of outcome studies comparing bona fide psychotherapies: Empirically, "All must have prizes." *Psychological Bulletin*, *122*, 203–215.

*White, J. (1998). Getting the learners' attention: A prototypical input enhancement study. In C. Doughty & J. Williams (Eds.), *Focus on form in classroom second language acquisition* (pp. 85–113). Cambridge: Cambridge University Press.

*White, L. (1991). Adverb placement in second language acquisition: Some effects of positive and negative evidence in the classroom. *Second Language Research*, *7*, 133–161.

**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]

Whittington, D. (1998). How well do researchers report their measures? An evaluation of measurement in published educational research. *Educational and Psychological Measurement*, *58*, 21–37.

Williams, J. (1995). Focus on form in communicative language teaching: Research findings and the classroom teacher. *TESOL Journal*, *4*, 12–16.

**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.

Wolf, F. (1986). *Meta-analysis: Quantitative methods for research synthesis*. Beverly Hills, CA: Sage.

Woods, A., Fletcher, P., & Hughes, A. (1986). *Statistics in language studies*. Cambridge: Cambridge University Press.

Wortman, P. (1994). Judging research quality. In H. Cooper & L. Hedges (Eds.), *The handbook of research synthesis* (pp. 97–110). New York: Russell Sage Foundation.

**Yang, L., & Givón, T. (1997). Benefits and drawbacks of controlled laboratory studies of second language acquisition. *Studies in Second Language Acquisition*, *19*, 173–194.

Yeaton, W., & Wortman, P. (1993). On the reliability of meta-analytic reviews: The role of intercoder agreement. *Evaluation Review*, *17*, 292–309.

*Yip, V. (1994). Grammatical consciousness-raising and learnability. In T. Odlin (Ed.), *Perspectives on pedagogical grammar* (pp. 123–139). Cambridge: Cambridge University Press.

Yirmiya, N., Erel, O., Shaked, M., & Solomonica-Levi, D. (1998). Meta-analyses comparing theory of mind abilities of individuals with autism, individuals with mental retardation, and normally developing individuals. *Psychological Bulletin*, *124*, 283–307.

Young-Scholten, M. (1999, March). *Focus on form and linguistic competence: Why Krashen is still right*. Paper presented at the Annual Conference of the American Association for Applied Linguistics, Stamford, CT.

*Zhou, Y.-P. (1992). The effect of explicit instruction on the acquisition of English grammatical structures by Chinese learners. In C. James & P. Garrett (Eds.), *Language awareness in the classroom* (pp. 254–277). London: Longman.

*Zobl, H. (1985). Grammars in search of input and intake. In S. Gass & C. Madden (Eds.), *Input in SLA* (pp. 329–344). Rowley, MA: Newbury House. [Two studies]

Appendix A

Summary of Several Substantive Features for Studies Included in Quantitative Meta-analysis

Study	Amount of instruction[a]	Dependent variables[b]	Independent variable conditions[c]	Instruction category[d]
Alanen 95	short	CCR, MJR	Rule (rule-oriented FonF)	FonF explicit
			Rule plus enhancement (rule-oriented FonF)	FonF explicit
			Enhancement	FonF implicit
			Reading for meaning (baseline)	FonM
Bouton 94	medium	(SR)	Explicit cognitive awareness (traditional explicit)	FonFs explicit
			Comparison group	No treatment
Cadierno 95	short	CCR, SR	Input processing	FonF explicit
			Traditional (traditional explicit)	FonFs explicit
			Control	No treatment
Carroll et al. 92	short	CCR	Corrective models	FonFs implicit
			Comparison group	No treatment
Carroll & Swain 93	short	MJR	Metalinguistic feedback	FonFs explicit
			Explicit correction: 'wrong'	FonFs explicit
			Corrective models	FonFs implicit
			Implicit correction: 'sure?'	FonFs implicit
			Control group	No treatment
Day & Shapson 91	long	(CCR), FR	Functional-analytic teaching (compound FonF)	FonF explicit
			Comparison group	No treatment

Study	Amount of instruction[a]	Dependent variables[b]	Independent variable conditions[c]	Instruction category[d]
de Graaff 97	long	CCR, MJR	Explicit: rule, enhancement, metalinguistic feedback (compound FonF)	FonF explicit
			Implicit: no rule, no enhancement, corrective models (baseline)	FonF implicit
DeKeyser 95	long	CCR, (MJR)	Explicit: rule (rule-oriented forms-focused)	FonFs explicit
			Implicit: no rule (baseline)	FonFs implicit
DeKeyser 97	long	CCR, SR	Same practice & test skill (traditional explicit)	FonFs explicit
			Reverse practice & test skill (traditional explicit)	FonFs explicit
			Both input and output practice (baseline)	FonFs explicit
DeKeyser & Sokalski 96-Study 1	short	CCR, SR	Input practice	FonFs explicit
			Output practice	FonFs explicit
			Rule-only (rule-oriented forms-focused)	FonFs explicit
DeKeyser & Sokalski 96-Study 2	short	CCR, SR	Input practice	FonFs explicit
			Output practice	FonFs explicit
			Rule-only (rule-oriented forms-focused)	FonFs explicit
Doughty 91	long	(CCR, FR, MJR)	Rule-oriented group (rule-oriented FonF)	FonF explicit
			Meaning-oriented group (enhancement)	FonF implicit
			Flood (baseline)	FonF implicit
Ellis, Rosszell, & Takashima 94-Study 2	brief	MJR	Garden Path: corrective model of induced error, followed by rule statement	FonFs explicit
			Modeling: rule explanation and pre-emptive models (traditional explicit)	FonFs explicit
Fotos & Ellis 91-Study 1	brief	MJR	Consciousness-raising	FonF explicit
			Teacher-fronted explanation (traditional explicit)	FonFs explicit
			Control	No treatment
Fotos & Ellis 91-Study 2	brief	MJR	Consciousness-raising	FonF explicit
			Teacher-fronted explanation (traditional explicit)	FonFs explicit
			Control	No treatment

Study	Amount of instruction[a]	Dependent variables[b]	Independent variable conditions[c]	Instruction category[d]
Harley 89	long	FR, SR	Functional-analytic teaching (compound FonF)	FonF explicit
			Comparison group	No treatment
Herron & Tomasello 88	brief	CCR	Feedback: clarification request, optionally followed by rule statement (baseline)	FonFs implicit (occasionally explicit)
			Modeling without rule explanation (pre-emptive model)	FonFs implicit
Hulstijn 89-Study 2	brief	CCR	Form group: anagram (form-experimental)	FonFs implicit
			Meaning group: affective reaction	FonM
			Form & meaning group: open-ended	FonF implicit
			Control	No treatment
Jourdenais et al. 95	brief	FR	Enhanced reading (enhacenment)	FonF implicit
			Unenhanced reading (baseline)	FonM
Kubota 94	brief	CCR, MJR	Metalinguistic feedback	FonFs explicit
			Explicit correction: 'wrong'	FonFs explicit
			Corrective models	FonFs implicit
			Implicit correction: 'sure?'	FonFs implicit
			Control	No treatment
Kubota 95a	short	CCR, MJR	Garden Path: corrective model of induced error, followed by rule statement	FonFs explicit
			Modeling: rule explanation followed by pre-emptive models	FonFs explicit
Kubota 95b	brief	CCR, SR	Consciousness-raising	FonF explicit
			Teacher-fronted explanation (traditional explicit)	FonFs explicit
			Control	No treatment

Study	Amount of instruction[a]	Dependent variables[b]	Independent variable conditions[c]	Instruction category[d]
Kubota 96	brief	CCR, MJR	Input plus metalinguistic feedback (metalinguistic feedback)	FonFs explicit
			Input plus corrective models (corrective models)	FonFs implicit
			Output plus metalinguistic feedback (metalinguistic feedback)	FonFs explicit
			Output plus corrective models (corrective models)	FonFs implicit
			Input practice only (baseline)	FonFs explicit
			Output practice only (baseline)	FonFs explicit
Leow 97	brief	CCR, SR	Problem-solving individual task with instruction to orient (metalinguistic task-essentialness)	FonF explicit
			Not aware subsample (baseline)	FonF explicit
Leow 98a	brief	CCR, SR	Teacher-fronted rule explanation (traditional explicit)	FonFs explicit
			Problem-solving individual task with instruction to orient (metalinguistic task essentialness)	FonF explicit
Leow 98b	brief	CCR, SR	Problem-solving individual task (baseline)	FonF implicit
			Problem solving individual task: form-inhibiting, with instruction to orient	FonF explicit
			Problem-solving individual task: form-essential, with instruction to orient (metalinguistic task-essentialness)	FonF explicit
			Problem-solving individual task: form-essential, without instruction to orient	FonF implicit
Long et al. 98-Study 1	brief	CCR	Recast: implicit corrective model, no rule	FonF implicit
			Model: pre-emptive model, no rule (other implicit)	FonF implicit

Study	Amount of instruction[a]	Dependent variables[b]	Independent variable conditions[c]	Instruction category[d]
Long et al. 98-Study 2	brief	CCR	Recast: implicit corrective model, no rule	FonF implicit
			Model: pre-emptive model, no rule (other implicit)	FonF implicit
Loschky 94	short	SR	Pre-modified input	FonF implicit
			Interactionally modified input (other implicit)	FonF implicit
			Unmodified input (baseline)	FonM implicit
Lyster 94	long	CCR, FR, SR	Functional-analytic teaching (compound FonF)	FonF explicit
			Comparison group	No treatment
Mackey & Philp 98	short	FR	Intensive recasts (recasts)	FonF implicit
			Unfocused interactionally modified input (baseline)	FonF implicit
Master 94	medium	SR	Explicit instruction: staged rule explanation plus practice and feedback (traditional explicit)	FonFs explicit
			Comparison group	No treatment
Nagata 93	medium	CCR	Complete rule metalinguistic feedback (metalinguistic feedback)	FonFs explicit
			Feedback locating general error source (baseline)	FonFs explicit
Nagata 95	medium	CCR	Complete rule metalinguistic feedback (metalinguistic feedback)	FonFs explicit
			Feedback locating general error source and location	FonFs explicit
Nagata 97a	medium	CCR	Complete rule metalinguistic feedback (metalinguistic feedback)	FonFs explicit
			Translation as feedback, without rule (baseline)	FonFs explicit
Nagata 97b	medium	CCR, FR	Complete rule metalinguistic feedback (metalinguistic feedback)	FonFs explicit
			Illustration with exemplars as feedback, without rule (baseline)	FonFs explicit

Study	Amount of instruction[a]	Dependent variables[b]	Independent variable conditions[c]	Instruction category[d]
Nagata 98	medium	CCR, SR	Input practice (baseline)	FonFs explicit
			Output practice	FonFs explicit
Robinson 96b	brief	MJR	Incidental group: reading comprehension (baseline)	FonM
			Implicit group: memory-based (form-experimental)	FonFs implicit
			Rule-search group: inductive explicit	FonFs explicit
			Instructed group: rule explanation (rule-oriented forms-focused)	FonFs explicit
Salaberry 97	short	(CCR, FR), SR	Input practice	FonFs explicit
			Output practice	FonFs explicit
			Control	No treatment
Scott 89	short	CCR, SR	Teacher-fronted grammar explanation (traditional explicit)	FonFs explicit
			Teacher-read flooded stories, no instruction to orient (baseline)	FonF implicit
Scott 90	short	(CCR, SR)	Teacher-fronted grammar explanation (traditional explicit)	FonFs explicit
			Teacher-read flooded stories, with instruction to orient (baseline)	FonF explicit
van Baalen 83	medium	FR	Explicit teaching: rule explanation and translation (traditional)	FonFs explicit
			Implicit teaching: no rule or translation, exemplars, audiolingual (traditional implicit)	FonFs implicit
			Compromise: both grammar-translation and exemplars (baseline)	FonFs explicit
VanPatten & Cadierno 93a	short	CCR, SR	Input processing	FonF explicit
			Traditional (traditional explicit)	FonFs explicit
			Control	No treatment

Study	Amount of instruction[a]	Dependent variables[b]	Independent variable conditions[c]	Instruction category[d]
VanPatten & Oikkenon 96	medium	CCR, SR	Input processing	FonF explicit
			Structured input: input processing without rule (other implicit)	FonF implicit
			Rule-only (rule-oriented forms-focused)	FonFs explicit
VanPatten & Sanz 95	short	CCR, FR, SR	Input processing	FonF explicit
			Control	No treatment
White et al. 91	long	CCR, (FR), MJR	Rule explanation, practice with metalinguistic tasks, & feedback (compound FonF)	FonF explicit
			Comparison group	No treatment
Williams & Evans 98	medium	CCR, MJR, SR	Flood group, no rule (flood)	FonF implicit
			Instructed group: rule, flood, and feedback (compound FonF)	FonF explicit
			Same tasks without flood (baseline)	FonM
Yang & Givón 97	long	(CCR, FR), MJR	Grammatical input group (other implicit)	FonF implicit
			Pidgin input group (baseline)	FonF implicit

[a]Brief treatment = less than 1 hour; short treatment = 1 to 2 hours; medium treatment = 3 to 6 hours; long treatment = 7 or more hours.

[b]CCR = constrained constructed response; FR = free response; MJR = metalinguistic judgment response; SR = selected response. Dependent variable types shown in parentheses were not available for independent analysis owing to insufficient reporting by primary researchers.

[c]Independent variable conditions are labeled according to instructional/experimental features reported across studies, sometimes departing from instructional labels used by the particular primary researchers.

[d]FonM = focus on meaning; FonF = focus on form, integration of form and meaning was sought; FonFs = focus on forms, integration of form and meaning was not sought or discussed; explicit = deduction (explicit rule presentation) or explicit induction (instructions to orient learner attention to forms or to induce metalinguistic hypotheses) was an element of the treatment; implicit = no explicit rule statement took place in the treatment; no instructions to attend to particular forms or to formulate metalinguistic hypothesis were given to learners. No treatment = group participated in pre- and post-tests only.

Appendix B

Publication Sources

Source	Study report frequency
Applied Language Learning	1
Applied Linguistics	7
Applied Psycholinguistics	2
CALICO	1
Canadian Modern Language Review	1
English Language Teaching Journal	1
European Journal of Cognitive Psychology	1
Foreign Language Annals	2
French Review	4
Institute for Research in Language Teaching Bulletin	2
Interlanguage Studies Bulletin Utrecht	1
JALT Journal	1
Language Learning	4
Language Learning and Technology	1
Modern Language Journal	8
Second Language Research	2
Studies in Second Language Acquisition	14
System	1
TESOL Quarterly	2
University of Pennsylvania Working Papers	2
Working Papers for Chofu Gakuen Women's Junior College	2
Book	1
Book chapter	16
Total study reports	77