ETS TOEFL

Quality Beyond Measure.

ISSN 1930-9317

TOEFL iBT™ Research Report

TOEFL iBT–14

The Effectiveness of Feedback for L1-English and L2-Writing Development: A Meta-Analysis

Douglas Biber

Tatiana Nekrasova

Brad Horn

February 2011

The Effectiveness of Feedback for L1-English and L2-Writing Development: A Meta-Analysis

Douglas Biber, Tatiana Nekrasova, and Brad Horn
Northern Arizona University

RR-11-05

ETS is an Equal Opportunity/Affirmative Action Employer.

As part of its educational and social mission and in fulfilling the organization's non-profit Charter and Bylaws, ETS has and continues to learn from and also to lead research that furthers educational and measurement research to advance quality and equity in education and assessment for all users of the organization's products and services.

Abstract

This research project undertook a review and synthesis of previous research on the effectiveness of feedback for individual writing development. The work plan was divided into two main phases. First, we surveyed all available studies that have investigated the effectiveness of writing feedback, including both quantitative and qualitative research, for students who have learned English as a first language (L1-English), students who have learned English as a second language (L2-English), and students who have learned second languages other than English. The results of this survey are described in a narrative overview of previous research pertaining to the role of feedback in the development of writing proficiency. The survey also identified the major theoretical constructs used in this research domain, providing the basis for subsequent statistical analysis.

Second, we built on this survey to carry out a meta-analysis of empirical studies in this research area. The goal of the meta-analysis was to provide a quantitative investigation of the extent and ways in which feedback has been effective, summarizing the findings of previous quantitative studies that have employed suitable statistical measures. Several analytical steps were required for the meta-analysis: developing a coding rubric; analyzing the research design and adequacy of reporting in studies to determine if they were suitable for inclusion; coding each study for all relevant research design factors; computing effect sizes for each study; and analyzing and interpreting the general patterns that hold across this set of studies.

The meta-analysis compared the gains in writing development with respect to several different kinds of feedback. Overall, feedback was found to result in gains in writing development. Beyond that, there were several predictable findings (e.g., that written feedback is more effective than oral feedback for writing development) and several other more noteworthy trends (e.g., that peer feedback is more effective than teacher feedback for L2-English students; commenting is more effective than error location; and in general, focus on form and content seems to be more effective than an exclusive focus on form).

Key words: feedback, writing development, meta analysis, commenting, error analysis

The Test of English as a Foreign Language™ (TOEFL®) was developed in 1963 by the National Council on the Testing of English as a Foreign Language. The Council was formed through the cooperative effort of more than 30 public and private organizations concerned with testing the English proficiency of nonnative speakers of the language applying for admission to institutions in the United States. In 1965, Educational Testing Service (ETS) and the College Board® assumed joint responsibility for the program. In 1973, a cooperative arrangement for the operation of the program was entered into by ETS, the College Board, and the Graduate Record Examinations® (GRE®) Board. The membership of the College Board is composed of schools, colleges, school systems, and educational associations; GRE Board members are associated with graduate education. The test is now wholly owned and operated by ETS.

ETS administers the TOEFL program under the general direction of a policy board that was established by, and is affiliated with, the sponsoring organizations. Members of the TOEFL Board (previously the Policy Council) represent the College Board, the GRE Board, and such institutions and agencies as graduate schools of business, two-year colleges, and nonprofit educational exchange agencies.

❖ ❖ ❖

Since its inception in 1963, the TOEFL has evolved from a paper-based test to a computer-based test and, in 2005, to an Internet-based test, TOEFL iBT™. One constant throughout this evolution has been a continuing program of research related to the TOEFL test. From 1977 to 2005, nearly 100 research and technical reports on the early versions of TOEFL were published. In 1997, a monograph series that laid the groundwork for the development of TOEFL iBT was launched. With the release of TOEFL iBT, a TOEFL iBT report series has been introduced.

Currently this research is carried out in consultation with the TOEFL Committee of Examiners. Its members include representatives of the TOEFL Board and distinguished English as a second language specialists from the academic community. The Committee advises the TOEFL program about research needs and, through the research subcommittee, solicits, reviews, and approves proposals for funding and reports for publication. Members of the Committee of Examiners serve four-year terms at the invitation of the Board; the chair of the committee serves on the Board.

Current (2010-2011) members of the TOEFL Committee of Examiners are:

Alister Cumming (Chair)	University of Toronto
Carol A. Chapelle	Iowa State University
Barbara Hoekje	Drexel University
Ari Huhta	University of Jyväskylä, Finland
John M. Norris	University of Hawaii at Manoa
James Purpura	Columbia University
Carsten Roever	University of Melbourne
Steve Ross	University of Maryland
Mikyuki Sasaki	Nagoya Gakuin University
Norbert Schmitt	University of Nottingham
Robert Schoonen	University of Amsterdam
Ling Shi	University of British Columbia

To obtain more information about the TOEFL programs and services, use one of the following:

E-mail: toefl@ets.org
Web site: www.ets.org/toefl

Acknowledgments

We would like to thank several anonymous reviewers for their comments on previous drafts of this report, and especially Yasuyo Sawaki for detailed comments and suggestions for revision. We also are appreciative to Brian Feliciano and Xiaoming Xi at ETS for their continuing help through all stages of the project.

Table of Contents

List of Tables

List of Figures

Page

1. Introduction

Feedback is generally regarded as essential for writing development at all levels, from students at the kindergarten through 12th grade (K–12) levels, to college freshman taking composition courses, to graduate students working on dissertation projects. Similarly, feedback has been considered essential for both first language (L1) and second language (L2) writing development.

Despite this widespread perception, much less agreement exists on the kinds of feedback that actually make a difference, or even on the kinds of gains in proficiency that can be expected from feedback. Numerous papers advocate one or another approach, and many other studies describe a writing course where a particular approach was used. Many other papers adopt a (quasi)experimental approach, measuring gains in writing proficiency that result from feedback.

Numerous factors must be considered in any study of feedback to determine which ones are actually influential. For example, feedback can be provided by the teacher, other students, or an automated system on a computer. Feedback can be written or spoken, and it can focus on content, organization, grammatical form, or usage (e.g., spelling). If written, feedback on form can comment on the existence of errors, identify the location of specific errors, or actually correct errors. And then, of course, questions must be addressed about how to measure potential improvements in writing performance resulting from feedback, for example, focusing on reduction in errors, the extent to which students incorporate revisions, or overall holistic assessments of writing quality.

Recently, Hyland and Hyland (2006) carried out a comprehensive survey of research on feedback, identifying several of the most important issues and describing numerous studies that investigate those issues (cf., DiPardo & Freedman, 1988). However, despite the large number of studies (over 200 in their survey), Hyland and Hyland concluded that there is surprisingly little consensus and most of the fundamental questions remain unanswered:

> While the research into feedback on L2 students' writing has increased dramatically in the last decade, it is clear that the questions posed at the beginning of this paper have not yet been completely answered. […] Nor are we a lot closer to understanding the long term effects of feedback on writing development. (p. 96)

In part, this lack of consensus results from the diverse research designs and methodologies used in previous studies of feedback. However, an additional limitation has been

the lack of quantitative techniques to document the state-of-the-art in this research domain. That is, previous survey articles, such as Hyland and Hyland (2006), have relied on descriptive narratives to survey previous research in this domain. However, those surveys provided no quantitative analysis of the distribution of research approaches and designs within the domain. For example, how many of these studies have been qualitative reports versus quantitative empirical studies? How many of these studies have used experimental designs versus other kinds of quantitative comparisons?

In fact, authors of state-of-the-art articles in applied linguistics usually pay little attention to the methods that they used themselves in carrying out the survey. That is, it has generally been assumed that the research for a survey article consists of finding as many publications on a topic as possible, determining the types of research and the main research issues represented by those studies, and then describing the studies that fall into each type. Such surveys rarely specify how articles were selected for inclusion in the review or provide any other evidence that the reader can use to evaluate the representativeness of the survey. Rather, the survey depends crucially on the expert knowledge of the authors. While such descriptions are a tremendous resource for future researchers beginning work in a particular domain, it is difficult to determine the extent to which the survey actually represents the research domain.

To address these concerns, recent research in applied linguistics has begun to advocate *systematic research syntheses*, applying the techniques of meta-analysis that have been developed over the past few decades for social science research. Systematic research syntheses differ from traditional literature surveys in three major ways (Norris & Ortega, 2006, pp. 6–7; see also Norris & Ortega, 2007, pp. 807–8):

1. The selection of studies to be included in the survey is a deliberate part of the research process, with explicit procedures for defining the population and identifying the research studies to be included or excluded from the survey.

2. Each research study is critically evaluated for the appropriateness of its research design and application of statistical procedures (rather than uncritically reporting study conclusions).

3. Each research study is analyzed with respect to the same set of design variables and values, applying a coding scheme developed for the entire meta-analysis.

A subset of the studies included in a systematic research survey will be suitable for a subsequent stage of analysis: a statistical meta-analysis based on comparison of effect sizes across studies. To be included in this stage of the research synthesis, a study must employ an experimental research design and be explicit and complete in its reporting standards. By comparing the magnitude of effect sizes across multiple studies in a research domain, it is possible to compare the importance of different factors based on the cumulative evidence of all empirical studies in the domain.

Two recent studies have applied the techniques of statistical meta-analysis to study writing development. These studies included some information on the effectiveness of feedback, although that was not the primary focus of either one. Truscott (2007) focuses on the quite restricted question of the extent to which error correction influences writing accuracy for L2-English students. This study concluded that overt error correction actually has a small negative influence on learners' abilities to write accurately. However, the meta-analysis was based on only six research studies, making it somewhat difficult to be confident about the generalizability of the findings.

The second study, Graham and Perin (2007) was much larger in scope but focused on writing instruction (for L1-English adolescent students) rather than the effectiveness of feedback. As a result, that study considered factors such as different instructional approaches (e.g., writing as product versus process); explicit instruction in grammar, sentence combining, writing strategies, and so on; prewriting activities; and the use of word-processing for writing. The only factor in that study that was directly relevant to the present inquiry was *peer assistance*, which was identified with a moderately large increase in writing quality.

The present report focuses exclusively on the influences of feedback for writing development, providing a large-scale and systematic synthesis of research on this topic. Because of the need to follow explicit procedures at all stages, the report is organized somewhat differently from a traditional literature review. In Section 2, we document the procedures that we used to describe the research domain and to attempt to construct an exhaustive catalog of research studies within that domain. We then describe our initial coding scheme, identifying the major ways in which studies of writing feedback can differ from one another. We also discuss the research designs and reporting standards that are required for a study to be suitable for inclusion in the statistical meta-analysis.

In Section 3, we provide an empirical survey of research studies in this domain, including discussion of the breakdown of studies across the major variables included in our coding scheme. Section 3 also describes the subset of studies that are appropriate for statistical meta-analysis of their effect sizes.

In Section 4, we turn to the procedures used for the statistical meta-analysis. Numerous analytical decisions are required for this stage of the synthesis, and our goal here is to describe those as fully and explicitly as possible.

Section 5 provides the most important information from this synthesis: the results of the statistical meta-analysis. In this section, we compare the magnitude (and dispersion) for the effect sizes of several different factors that have been hypothesized to influence the effectiveness of feedback for writing development. Based on these analyses, we are able to provide an overall perspective on the influence of feedback, identifying factors that seem to make a difference for writing development and those that seem to be less influential.

A summary and discussion of the statistical meta-analysis is taken up in Section 6.

2. Procedures I: Describing the Research Domain

The first major stage of this project was to describe the research domain. Research for this stage was carried out in three steps: First, we conducted a literature search to identify all relevant studies, employing the procedures described in Section 2.1. Second, we developed an explicit coding scheme, described in Section 2.2, that included all major variables represented in the research designs of these studies and the major values that were distinguished for those variables. Finally, we coded each research study for all variables in our coding scheme.

2.1. The Literature Search

The literature search began with an operational definition of the population of studies, followed by a comprehensive sampling of studies in that population. For the purposes of this search, we attempted to identify all studies that addressed the central research question of our research synthesis:

Which kinds of feedback are influential for which kinds of gains in writing proficiency?

This research question has two main components: (a) the different operationalizations of feedback (kinds of feedback) and (b) the range of outcome measures (kinds of gains in writing

proficiency). Thus, we included articles that investigated different sources of feedback (e.g., teacher, peer, computer), as well as different forms of feedback (e.g., direct correction, editing codes, highlighting) delivered in different modes (i.e., spoken, written, and computer mediated). For similar reasons, in addition to articles that report development in terms of writing proficiency measures, we also included articles that reported results from other outcome measures (e.g., surveys of student attitudes, analyses of post-feedback revisions).

We included studies of both native and/or nonnative English speaking students (including developing and remedial writers). Our goal in doing this was to allow comparison of the two populations, asking whether feedback is influential in the same ways and to the same extent in L1 and L2 populations.

Location and selection of research studies. The first step in our survey was to identify research journals that could potentially publish articles on feedback. This was done by exploring library catalogs and databases and by including any journal cited in previous survey studies. The following journals were included in this step:

Academic Writing Across the Disciplines
Applied Linguistics
Assessing Writing
Australian Journal of Language and Literacy
British Journal of Educational Technology
CALICO Journal
CALL Electronic Journal
Canadian Modern Language Review
College Composition and Communication
Computer Assisted Language Learning
Computers and Composition
Computers & Education
ELT Journal
English for Specific Purposes
English Journal
Foreign Language Annals
International Review of Applied Linguistics
Issues in Writing
Journal of Basic Writing
Journal of Educational Psychology
Journal of Educational Research
Journal of English for Academic Purposes
Journal of Second Language Studies
Journal of Second Language Writing
Jnl of Technical Writing & Communication
Language Learning
Language & Learning Across the Disciplines
Language Teaching
Language Teaching Research
Modern Language Journal
ReCALL
Research in the Teaching of English
RELC Journal
Rhetoric Review

Second Language Research
Spaan Fellow Working Papers
Studies in Second Language Acquisition
System
Teaching English in the Two Year College
TESL Canada Journal
TESL-EJ
TESOL Journal
TESOL Quarterly
Writing Center Journal
Written Communication

For each of these journals, we searched the online table of contents to identify all articles that had any of the following keywords: feedback, response, comment(ing), revision, peer, and writing. The range of dates searched was dictated by the archival status of individual journals but in general spanned the period 1980–2007. In addition, an online search of the ERIC database was conducted using the keywords *writing* and *feedback*. Our literature search focused primarily on studies published in academic journals. Further, as individual articles were being analyzed, the list of references in each was reviewed to identify additional articles (including studies published in edited books) that had not yet been collected. The studies included in our literature survey are mostly published research articles; we made no systematic attempt to include studies from the "fugitive" literature (e.g., unpublished papers, dissertations, conference presentations), apart from research papers identified through the ERIC database.

Using these methods, we were able to collect articles representing a variety of epistemological traditions. Unlike the methods used for some other meta-analyses, we did not adopt *a priori* exclusion criteria regarding research methodology (e.g., accepting only experimental or quasi-experimental studies). Instead, the articles included in our survey ranged from tightly controlled experimental studies to qualitative case studies. This inclusive approach allowed us to evaluate the maturity of the research domain before selecting empirical studies for the quantitative meta-analysis.

While articles were not excluded from the survey on the basis of research methodology, we did exclude studies that were not in the research domain of focus here. In particular, we excluded the following:

- studies focusing on oral (rather than written) production;
- studies in which computer-mediated chat was the target of feedback (because engaging in chat is a different communicative enterprise from the writing tasks normally considered in studies of writing development). (Note: we did include studies

that investigated chat as the means through which feedback on writing was delivered.); and

- studies focusing on the writing of special-needs student populations (e.g., deaf students).

2.2. Identifying the Parameters of Variation Among Research Designs and Coding the Study Reports

The central research question motivating this research synthesis has two main components: the different kinds of feedback and the different measures of improvement in writing proficiency. We thus began this project by carrying out preliminary research on how these two constructs have been approached in previous research.

Then, with that background, we developed an explicit coding rubric. The goals of this step were to identify all important factors that varied across feedback studies (e.g., age of the subjects, type of writing task required, type of feedback provided) and to itemize the possible values for each of those variables. This rubric was developed inductively, by reading through a wide sample of research studies to identify various ways in which their research designs could vary. The rubric was subsequently applied for an empirical description of this research domain (described in Section 3).

Operationalizations of feedback in the research literature. On initial consideration, feedback might seem to be a simple construct—providing a constructive evaluation of writing quality to the student. However, in actual practice and in the research literature, an extremely wide range of variation was found in the actual realization of feedback. These differences can be described with respect to five variables: type, focus, tone, mode, and source.

Type of feedback. In research on traditional teacher-generated feedback, the distinction between *direct* and *indirect feedback* has been one focus of studies in the areas of writing and second language acquisition (SLA) research (e.g., Ferris, 2003, 2006; Ferris & Roberts, 2001; Robb, Ross, & Shortreed, 1986). The term *direct feedback* is used to denote instances where the writing instructor makes an explicit correction to the student's text (e.g., by writing in the correct grammatical form), while *indirect feedback* denotes instances where the instructor indicates that something about the student's writing is problematic (e.g., by underlining an ungrammatical construction and/or marking the problematic section of text with a special code) but does not

provide an immediate correction. In actual practice, direct feedback is rarely used as a treatment in empirical research, while numerous types of indirect feedback have been investigated. These include identifying the location of problems, providing comments in the margins, global comments at the end of a paper, and even oral comments given to the student.

Focus of feedback. This area of research has dealt with the features of student writing (e.g., lexis, grammar, mechanics, organization, content) that the feedback provider chooses to focus on. As noted above, much feedback research has focused on error correction. Researchers on second language writing research distinguish between grammatical and word choice errors, because such "L2 errors" are thought to stigmatize L2 users. For example, Ferris (1999) divided such errors into two classes, which she labeled *treatable* and *untreatable*. Treatable errors are those that can be addressed through explicit instruction and include language features such as article usage and subject-verb agreement (i.e., rule-governed constructions). Untreatable errors are those that are less readily teachable in that they are not governed by a clear or simple set of rules. Problems with word choice are one example Ferris gave of such untreatable errors.

The predominating emphasis on error correction seems to be motivated by the perceived severity of different error types among readers of L2 texts. However, not all teacher comments address aspects of student language use that can be objectively characterized as incorrect or even problematic (e.g., positive feedback, clarification questions). Furthermore, many student writers desire guidance in these additional areas, especially as they reach more advanced levels of writing proficiency (Leki, 2006). While feedback on surface level errors may be comparably easy to provide (both for human teachers and computer programmers), an important question is whether this type of feedback leads to greater gains in student writing proficiency than more holistically focused feedback on text content, organization, or audience/purpose.

Tone of feedback. Following from the idea that not all feedback focuses on student errors, it is also the case that feedback can vary in the degree to which it praises areas of strength or criticizes areas of weakness (see, e.g., Hyland & Hyland, 2001). Concern has been expressed in the literature that overly negative feedback will adversely affect the student's motivation. At the same time, it is possible that some students may view positive feedback as less useful than critical feedback that identifies features of their writing that need to be revised. Thus, an important concern for instructors is determining the best tone for constructive criticism, given (a)

the nature/amount of feedback that needs to be provided and (b) the nature of the teacher-student interpersonal relationship.

Mode and source of feedback. Finally, feedback can be provided through any available channel, or *mode*: oral, written, or computer mediated. Although it has not been a major factor in previous research, several studies considered the influence of one mode of delivery over another. Similarly, feedback can be provided by the teacher or by other students, or even generated automatically by computer.

Operationalizations of writing development: Outcome measures of the effects of feedback. To demonstrate the effectiveness of feedback, researchers have used measures of writing proficiency (e.g., Chandler, 2003; Min, 2006), as well as survey instruments designed to elicit student perspectives (e.g., Ferris, 1995). Writing proficiency measures that have been used in feedback research included ratings obtained from classroom teachers and/or trained judges using holistic and analytic rating scales, as well as other measures of syntactic and lexical complexity. Student perspectives or changes in student attitudes have been elicited using both qualitative approaches (e.g., interviews) and quantitative instruments (e.g., surveys). A third approach to analyzing the effectiveness of feedback has been to tabulate the number of suggested revisions that were ultimately adopted by the student in subsequent drafts (e.g., Min, 2006).

The coding rubric. The first major stage of a systematic research synthesis is to undertake an empirical analysis of the research domain, documenting the ways in which the central research questions have been approached within that domain. This description is also required to evaluate whether the research domain is sufficiently mature to permit a statistical meta-analysis. In the present case, our preliminary reading indicated that much of the research on writing for the past two decades has eschewed quantitative methods in favor of more qualitative approaches, especially in the area of first language composition research. While qualitative work adds to our collective understanding of how students develop their writing skills, such studies cannot be included in a quantitative meta-analysis. Thus, the ultimate goal of the analysis in this stage is to determine whether enough experimental studies—with clearly documented research designs and statistical results—exist to permit the application of meta-analytic techniques.

To accomplish the empirical analysis of the research domain, it is first necessary to develop a coding rubric that itemizes all important factors that vary across feedback studies, as well as all possible values for each of those variables. This rubric is developed inductively, by

reading through a wide sample of research studies to identify various ways in which their research designs could vary.

The coding rubric developed for the present project includes 16 variables:

- Research paradigm
- Statistical analysis
- Design variables
- Target language
- Proficiency level (for L2 studies only)
- Number of student participants
- Age/grade level of student participants
- Genre of writing task(s)
- Length of writing task(s)
- Source of feedback
- Mode of feedback
- Focus of feedback
- Tone of feedback
- Type of feedback
- Outcome measures
- Specific focus for outcome measures of writing proficiency.

These variables were used to categorize the types of research studies in this research domain, and thus it was necessary to develop an exhaustive list of values for each variable. These values and variables are shown in Table 1 (with the codes used for the meta-analysis given in square brackets; see Appendices B and C).

Table 1

The Coding Rubric: Variables and Values for Each Variable

Variables	Values	Further details
Research paradigm	Quantitative	Experimental; quasi-experimental; correlational; survey
	Qualitative	Ethnographic; case study; interviews
	Mixed methods	Combination of quantitative measures and qualitative description
	Thought piece	Theoretical argument; pedagogical primer; no original empirical data
Statistical analysis	Statistical tests reported	Record statistic(s) used, including descriptive statistics
	Statistical tests not reported	
Design type	Intact group(s)	
	Random group assignment	
	One group	
	Treatment/control [TC]	
	Pretest/posttest [PP]	
	Posttest only	
	Descriptive/ex-post facto	
Target language	L1 English [E1]	English composition studies where no mention of nonnative speaker (NNS) participants is made
	L2 English [E2]	Most North American L2 writing studies
	Mixed L1 & L2 [MX]	Comprises native speaker (NS) & NNS students
	L1 Other [O1]	Students whose native language is not English, learning to write in their native language (e.g., a study of the composing processes of Dutch L1 children)

Variables	Values	Further details
	L2 Other [O2]	Students whose native language is English, learning to write in a second language (e.g., a study of U.S. college students learning L2 Spanish composition)
Proficiency level (for L2 studies only)	Proficiency level reported	[L] = Low; [H] = High
	Proficiency level not reported	
Number of student participants	*N*-size reported	Number of participants reported in study
	N-size not reported	
Age/grade level of student participants	Age or grade level reported	
	Age not reported	
Genre of writing Task	Correspondence	Business letters; personal letters; email; memos; faxes
	Creative	Fiction; poetry
	Pedagogical [PD]	"Learning" genres, such as five-paragraph essays
	Personal	Diaries; journals; reflective essays
	Research/academic	Scientific research articles; dissertations; theses; term papers
	Other genres [O]	Legal writing; journalism
	Genre not reported	
Length of writing Task	Length reported	
	Length not reported	
Source of feedback	Teacher [TE]	Feedback from course instructor
	Peer [PE]	Feedback from another student
	Tutor	Feedback from writing center tutor

Variables	Values	Further details
	Student	Self-correction
	Computer	Computer-generated feedback (not just computer-*mediated* feedback)
	Other [O]	
	Source not reported	
Mode of feedback	Oral [OR]	Face-to-face conferencing; tape-recorded comments
	Written [WR]	Marginal comments; end comments; editing codes; circles/underlines
	Computer-mediated [CM]	Internet chat; email
	Mode not reported	
Focus of feedback	Grammar	Subject-verb agreement errors; tense/aspect errors; pedagogical grammar issues in L1 studies
	Vocabulary	Collocation errors; other word choice issues
	Spelling	Spelling errors
	Organization [O]	Topic sentence; discourse markers; transitions; paragraphing; conclusion; *order* of content
	Content [C]	*Correctness* of content; *completeness* of content
	Punctuation / mechanics	Comma errors; end punctuation errors; indentation; capitalization; but *not* spelling
	Other	Anything not captured by the other values for this variable
	Form [F]	Grammar, spelling, punctuation
	Content and form [C,F]	Content and form
	Focus not reported / specified	
Tone of feedback	Negative	Comments on what the student has done wrong

Variables	Values	Further details
	Positive	Comments on what the student has done right
	Mixed	Comments on both strengths and weaknesses of text
	Tone not reported/specified	
Type of feedback	Location of error/problem/issue indicated [LO]	Location of an error is marked (circled, underlined), but no feedback is given on why it is an error or how it might be corrected
	Comment [CM]	Teacher/peer writes prose comments in the margin or at the end of the paper
	Other [O]	Other types of feedback, including direct correction/reformulation [DC]; editing codes [EC], error existence [EX], metalingusitic explanation of an error [ML], spoken explicit comments [SE], spoken implicit comments [SI]
	Multiple [M]	Multiple types of feedback are provided, such as both location and explanatory comments
Outcome measures	Writing proficiency measures [WP]	Holistic ratings of writing quality, measures of spelling accuracy, grammatical accuracy
	Attitude measures	Likert-scale items
	Records of composition strategies/processes employed	Records of time spent planning, drafting, etc.; eye-tracking records
	Records of revisions [RV]	Number/extent of revisions made
	Other [O]	
Focus for outcome measures of writing proficiency	Grammar [GR]	
	Spelling [SP]	
	Holistic [H]	
	Content [C]	

Most studies included in our study involved revisions made to an essay based on the same prompt over a period of time in response to different kinds of feedback. (McGroarty and Zhu 1997 was exceptional in this regard, because they evaluated writing development across essays based on different prompts.) The outcome measure for most quantitative studies was a measure of writing proficiency (either holistic quality or grammatical accuracy) based on evaluation of the final (revised) written product. However, in a few cases, studies simply documented the extent to which a student made any revisions, regardless of the contributions those revisions made to the quality of the final product.

Coding the studies. The initial coding of studies for general variables, such as the research approach, general design type, and target population, was carried out by the second and third authors of the report (TN and BH). Any controversial coding decisions were discussed by all three authors and resolved through consensus.

Subsequently, more detailed coding was undertaken by the second author (TN) for the purposes of the quantitative meta-analysis. The first step for this process was to identify the sub-set of studies that were suitable for inclusion in the analysis: studies that were published in the last 25 years, used quantitative measures, had an experimental (or at least quasi-experimental) design, were explicit about the types of feedback that were provided, employed a clear basis for comparison, and included an outcome indicator that measured change in students' writing proficiency or behavior (see Section 4.1 below). Any controversial coding decisions during this process were resolved through consensus by discussion between the first two authors (DB and TN).

3. Empirical Survey of the Research Domain

Based on the sampling methods described in Section 2.1, we collected a total of 306 articles that addressed the effectiveness of feedback for writing development. Our goal here was to obtain an exhaustive sample of studies published in the last 30 years, resulting in a much larger collection of publications than in some previous meta-analyses.

Figure 1 shows the breakdown of studies across year of publication. The trend here shows a dramatic increase in the number of feedback studies over the past 30 years. This trend reflects two factors. First is the general information explosion, with an increase in the number of academic journals and publications in all disciplines, and the more specific increase in the number of studies investigating the effect of feedback. Second, and more important for us here, is that this increase suggests researchers (and teachers) have shifted away from an uncritical

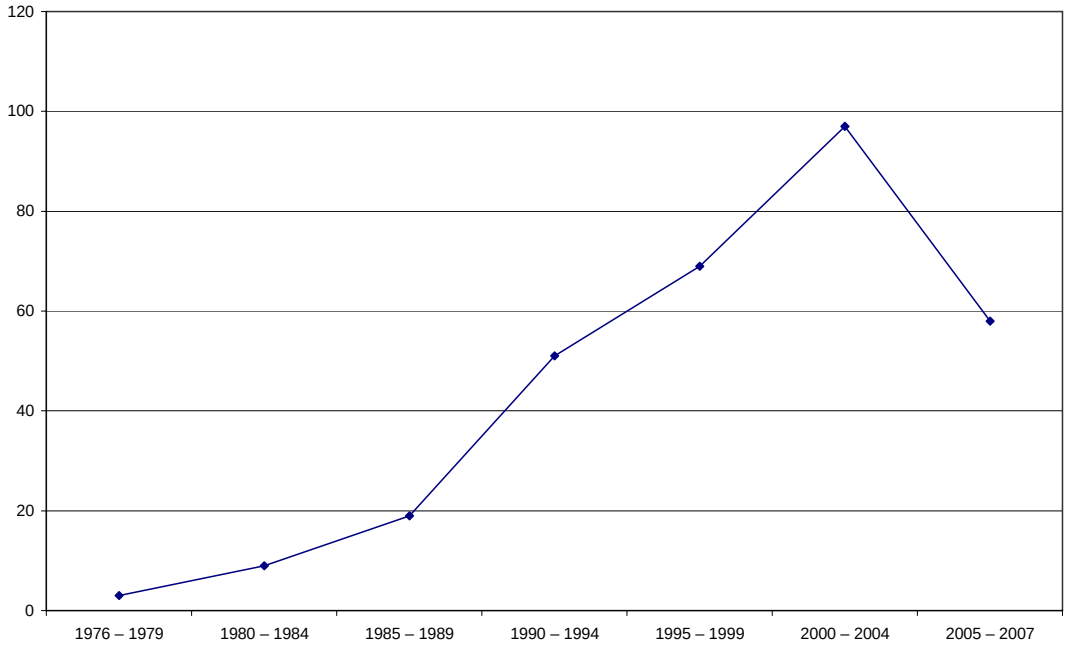

Figure 1. **Number of publications by date.**

belief in the effectiveness of feedback toward a recognition that feedback can take many different shapes and its effectiveness needs to be studied in its own right. (The last period includes only 2.5 years, accounting for the apparent decrease in publications.)

Figure 2 shows that studies in this domain have adopted the full range of research methodologies, with both qualitative and quantitative approaches represented by a large number of studies. In addition, numerous *thought pieces*—either survey articles describing previous research on feedback or general discussion articles—are included.

Figure 3 shows that the relative preference for one or another research approach has remained relatively constant across time, with quantitative studies being slightly more common than qualitative studies. The one notable departure from this pattern is in the period 2000–2004, which showed a dramatic increase in the number of qualitative studies while the number of quantitative studies remained constant. This shift might reflect a more general paradigm shift influenced by postmodern thinking in general, valuing ethnographic reports of individual case studies over reports of the general trends in a large sample of individuals. Because comparatively few studies are included in the most recent period, it is not clear whether this trend continues.

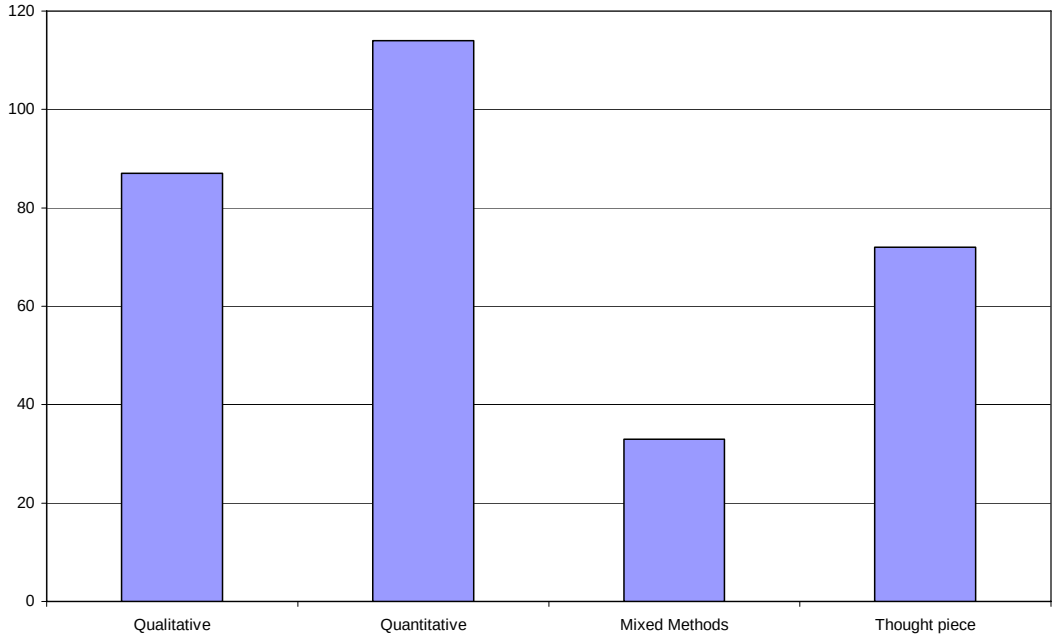

Figure 2. **Number of research publications by research approach.**

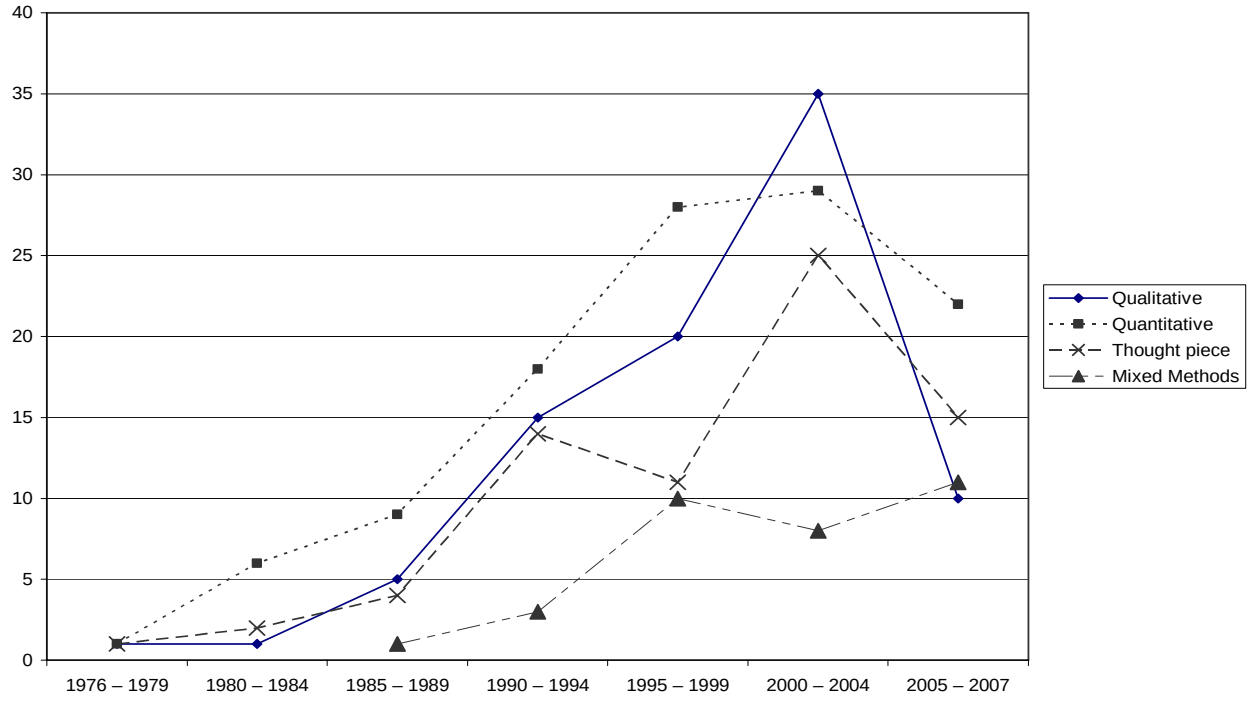

Figure 3. **Number of publications from each research approach by date.**

Studies have also varied in the target population that has been investigated (although many studies do not provide full details on the subjects). For example, Figure 4 shows that learners of English have been the primary target of investigation, although there have also been numerous studies of feedback that focused on the writing development of native English speakers. However, there has been a shift in research focus across time, as shown in Figure 5: Through the 1980s, equal interest was found in the influence of feedback for both L1 and L2 learners of English (although the number of studies is comparatively few). However, by the mid-1990s, a dramatic shift in focus occurs with many more studies focusing on learners of English than on the writing development of native English speakers.

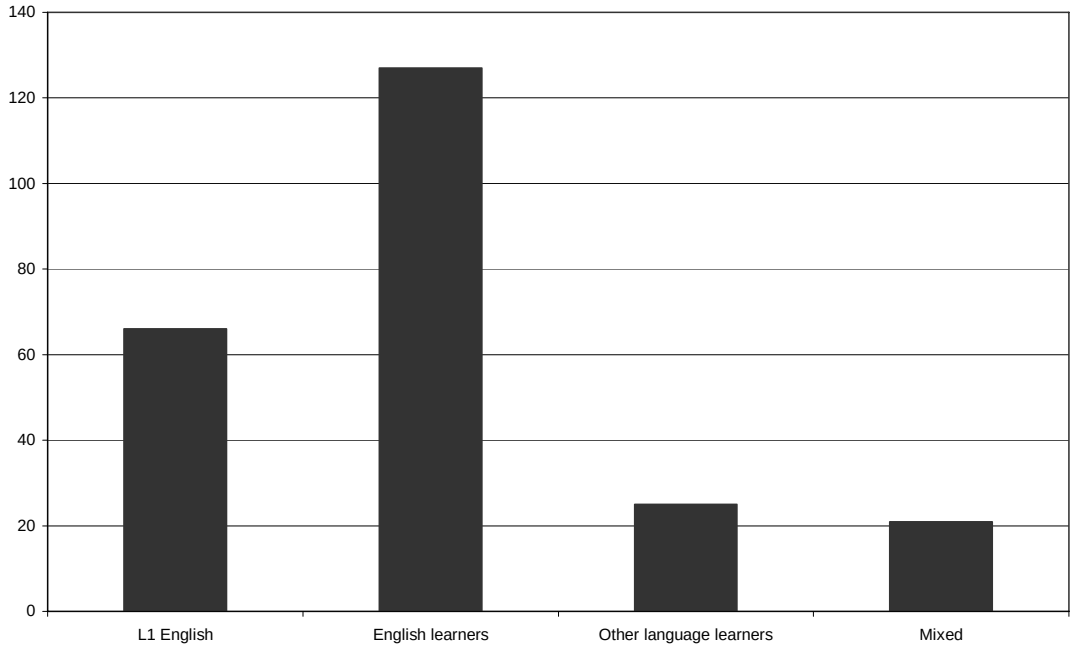

Figure 4. **Number of publications by target population**

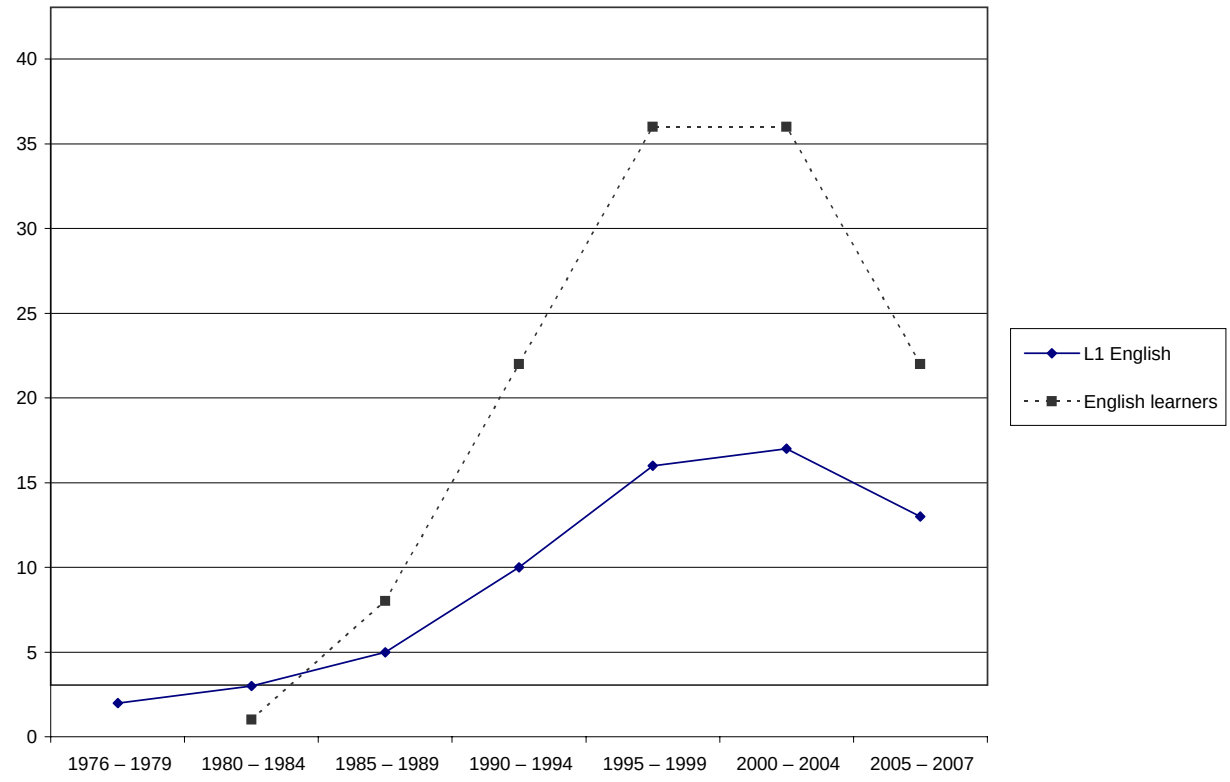

Figure 5. **Number of publications focusing on L1 versus L2 learners by date.**

Feedback studies that were focused on learners of English investigated the full range of proficiency levels, as shown in Table 2.

Table 2

Breakdown of Studies by Proficiency Level of the Target Population (Includes Only Studies of English Learners)

Proficiency levels	Number of studies
Low/beginner	17
Intermediate	19
Advanced	26
Mixed	21
Unspecified	44
Total	122

Table 3 shows that the majority of feedback studies (whether L1 or L2) have focused on the writing of college-aged students, while comparatively few studies have investigated the influence of feedback for younger students.

Turning to the nature of the student writing, Table 4 shows that the overwhelming majority of feedback studies have used pedagogical writing tasks, such as the five-paragraph essay.

Table 3

Breakdown of Studies by Age of the Target Population

Target population age	Number of studies
Ages 4–9	8
Ages 10–12	10
Ages 13–18	15
College-age	159
Other adult ages	29
Unspecified	85
Total	306

Table 4

Breakdown of Studies by the Genre of the Writing Task

Genre of writing task	Number of studies
Pedagogical	187
Personal correspondence	7
Personal journal or diary	7
Creative writing	2
Research/academic writing	8
Other/unspecified	95
Total	306

Since the central research question for this project focuses on the effectiveness of feedback, we coded five different variables to capture the different ways in which feedback was realized: source, mode, focus, type, and tone. As Table 5 shows, the large majority of these studies focus on feedback given by the teacher. In this regard, feedback studies probably reflected typical classroom practice, but they were at odds with many theoretical discussions that advocated the utility of peer feedback.

Most feedback on student writing was communicated in writing, either using a computer (32 studies) or with feedback written by hand (94 studies). Many studies did not report the mode of feedback; only 37 studies reported giving oral feedback, and an additional 36 studies used multiple modes.

Most of the studies that reported on the focus of feedback compared the influence of multiple categories (80 studies—usually a focus on both form and content). However, most studies did not report a specific focus, while only 14 studies had a single focus: 3 on content, 9 on grammatical form, and 2 on spelling.

Similar to the incomplete reporting typical of the other parameters, most studies in our sample did not report the particular type of feedback. For the remaining studies, Table 6 shows that written comments are the most common type of feedback, while a large number are also based on multiple types of feedback.

Table 5

Breakdown of Studies by the Source of Feedback

Source of feedback	Number of studies
Computer	18
Peer/other students	34
Self criticism	17
Peers + self	6
Teacher	119
Teacher + other	38
Other/unspecified	74
Total	306

Table 6

Breakdown of Studies by the Type of Feedback

Type of feedback	Number of studies
Comments	66
Error code	2
Direct correction of error	11
Location of error	4
Multiple types	35
Other/unspecified	188
Total	306

Only 17 of the 306 studies in our sample reported on the tone of feedback. Of those, 15 claimed to provide both positive and negative feedback, and 2 provided only positive feedback. It seems unlikely that this emphasis on positive feedback was equally typical for the 289 studies that did not report on tone.

Finally, we noted in Section 2.1 above that the central research question motivating this research synthesis has two main components: the kinds of feedback (described in the preceding paragraphs) and the resulting gains in writing proficiency. Table 7 shows that there is very little agreement on the best way to operationalize writing proficiency or development. 102 of the studies in our sample provided no specific measure of writing development. The remaining 204 studies, though, used a wide range of different measures, including questionnaires to determine student attitudes, direct comments on progress from teachers or students, and a record of the extent to which essays have been revised. Surprisingly few studies included a direct measure of writing quality, which might include scores for grammatical accuracy, content, organization, or an overall holistic rating for quality.

In sum, our survey of research relating to feedback on writing shows that considerable depth exists in this research domain, with numerous studies undertaken from multiple perspectives. About half the studies in this research domain have been quantitative, and those studies have included many variants of research design. There are advantages to this diversity, in

Table 7

Breakdown of Studies by the Outcome Measures of Writing Development

Outcomes used to measure writing development	Number of studies
Attitude measures	35
Revisions	39
Attitude measures plus revisions	12
Composition strategies	7
Composition strategies plus revisions	3
Essay score for quality or grammatical accuracy	40
Essay score plus attitude measures	18
Essay score plus revisions	18
Teacher or student comments on progress	32
Other/unspecified	102
Total	306

that each new research study considers slightly different research questions from preceding studies. For the purposes of a quantitative meta-analysis, however, this diversity, which depends on the existence of multiple studies that are directly comparable, also presents disadvantages.

The following section turns to the methods of meta-analysis and an evaluation of this research domain to determine if it is suitable for this approach.

4. Procedures II: The Quantitative Meta-Analysis

The meta-analysis proceeded in three major steps:

1. All publications in the larger sample were examined to identify the set of studies that were suitable for inclusion in a meta-analysis.
2. Effect sizes were computed for the outcome variables in each of those studies.
3. Mean effect sizes were computed for each treatment variable as the basis for determining the influence of different forms of feedback on writing development.

We describe each of these methodological steps in turn in the following subsections.

4.1. Identifying the Subset of Studies That Are Suitable for Meta-Analysis

During the coding of research articles described in Sections 2 and 3, we made an initial determination of whether a study was potentially suitable for inclusion in the quantitative meta-analysis. There were four major requirements for this initial screening (following the procedures used in Norris & Ortega, 2000, pp. 432–33):

1. The study was published in the last 25 years (between 1982 and 2007).
2. The study used quantitative measures and had an experimental (or at least quasi-experimental) design. Specifically, the study had to use and report on quantitative measures of effectiveness, for specific types of feedback.
3. The independent variables measured feedback characteristics, including source of feedback (e.g., teacher, peer, tutor, student, computer), focus of feedback (e.g., grammar, vocabulary, spelling, organization, content, mechanics, rhetorical organization), or type of feedback (direct comment, editing code, rating, etc.).
4. The dependent variables included an outcome indicator that measured the impact of specific types of feedback on participants' writing behavior, including writing proficiency (e.g., grammatical accuracy or holistic quality rating), increase in text length, attitude, strategies/processes employed, number/extent of revisions made.

Based on these criteria, 112 studies were identified as potentially relevant for the meta-analysis. These studies were then subjected to a second round of closer scrutiny to determine if the design and reporting standards were in fact adequate for our purposes here. Unfortunately, it turned out that a large number of additional studies were excluded in this second phase, for the following reasons:

1. The research design was not suitable for inclusion. That is, following Norris and Ortega (2000), we included only studies with designs based on mean differences: a pretest/posttest design, or a control group/experimental group design.

   Several studies were excluded because they used correlational designs. Although it is possible to compute effect sizes from such designs, these studies addressed fundamentally different kinds of research questions, and so they could not be readily compared to the effect sizes from group-comparison studies. Twenty-four studies were excluded for this reason.

2. The study addressed a different research question from the one that we are investigating in our project (e.g., studies on whether males/females produced more errors or studies on whether grading rubrics are biased to favor males or females). Some of these studies had a pre-post test design, but no actual feedback was provided. Sixteen studies were excluded for this reason.

3. The study was incomplete in its reporting of the design, sample, or statistical findings. Specifically, to be included in the meta-analysis, the study must report one of the following: (a) the sample size, mean scores, and standard deviations for each group, (b) between-groups *t* or *F* values together with *df*, or (c) individual scores on outcome measures for all participants. Twenty-four studies failed to meet these reporting standards for statistical tests (e.g., reporting only significance with no *df* or no *t* value, or reporting mean scores with no standard deviation); these studies were thus excluded.

4. The study provided no clear basis for comparison. These were mostly studies of a single group that reported only posttest results. Fourteen studies were excluded for this reason.

5. The study compared multiple treatment groups with respect to a single posttest with no pretest and no control group. For example, one group received feedback on content, while a second group received feedback on form; or one group received direct correction of errors, while only general comments were provided to a second group. Although these studies addressed some of the central research issues of our synthesis, they could not be included in the meta-analysis because it was not possible to isolate the influence of individual factors. As Norris and Ortega (2006) noted: "Direct comparisons between treatment conditions are not made, because they would be idiosyncratic to the particular study, and therefore not comparable with other studies that did not operationalize the same two treatments" (pp. 27–28). Eleven studies were excluded for this reason.

In sum, 89 additional research studies were excluded at this stage, leaving only 23 published papers (reporting on 25 different studies) that were directly comparable and otherwise suitable for

inclusion in the meta-analysis. At this point, the large majority of studies in this research domain were noted as not suitable for inclusion in a meta-analysis for three general reasons:

1. Many studies in this domain were qualitative (and often anecdotal), or thought pieces, based on researchers' observations and perceptions.
2. Many of the quantitative studies were not carefully designed, or the reporting standards were not adequate for the purposes of meta-analysis.
3. Several studies were carefully designed and implemented, but they simply addressed different research questions from the one this study focuses on.

Thus, although we were able to identify a large number of research studies in our initial survey (306 studies), relatively few of these could be used in the subsequent meta-analysis (only 25 studies).[1]

4.2. Computing Effect Sizes for the Outcome Variables

The second step in the meta-analysis was to compute an effect size for each outcome variable that reflects the influence of feedback. Again following Norris and Ortega (2000, 2006), Cohen's *d*-index was selected as the most appropriate effect size estimate and calculated for each finding related to feedback that was reported with sufficient data. Cohen's *d* represents the size or importance of a difference, either between a treatment group and a control group, or between a pretest and a posttest. (Correlational designs were not included in the final meta-analysis because they are not comparable to group comparisons designs.) In either case, this difference is interpreted as reflecting the influence of some treatment. Cohen's *d* is essentially a kind of standard score representing standard deviation units. It is calculated for a specific outcome measure by subtracting the mean scores for the two groups and then dividing this difference by the pooled standard deviation of the two groups. (There are numerous reference works that provide specific formulae to be used for the computation of effect size from different primary statistics; see e.g., Cohen, 1988; Lipsey & Wilson, 2001; Norris & Ortega, 2000; Rosenthal, Rosnow, & Rubin, 2000).

The studies included in our meta-analysis are about evenly split between studies with treatment-control designs and studies based on comparison of a pretest and posttest given to a single group (i.e., with no control group; see Section 5.1 below). Treatment-control designs (independent samples) are much stronger, allowing the researcher to isolate the influence of

feedback (the treatment) apart from other factors. In contrast, pretest versus posttest (dependent samples) designs that include only a single group are relatively weak because there is no control for the influence of natural development that occurs over the course of the study (see Section 5.2 below). For this reason, our analyses in the following sections distinguish between these two design types to the extent that it was feasible, reporting separate mean effect sizes for each type. In general, the results are consistent across both treatment-control and pretest-posttest designs, but the results for the latter should be interpreted with caution.

The computation of effect size also differs for the two design types (although both are referred to as Cohen's *d*). For studies that employed a treatment-control design, we used an online calculator to compute effect sizes (Becker, 1999):

`http://web.uccs.edu/`lbecker`/Psy590/escalc3.htm`

This calculator uses the following standard formula (which is consistent with Cohen's (1988, p. 44) formula):

$$\text{Cohen's } d = (M_1 - M_2)/ \sigma_{pooled}$$

$$\text{where } \sigma_{pooled} = \sqrt{[(\sigma_1^2 + \sigma_2^2) / 2]}$$

The reporting standards for all treatment-control studies included in our sample were high, so all studies included the descriptive statistics required for the formula (i.e., the mean scores and standard deviations for each group).

Studies that employ dependent-sample designs (i.e., pretest/posttest designs) are more problematic. In theory, the original individual scores for each subject should be used to compute effect sizes for this type of design. That is, the appropriate formula for computing the pooled standard deviation for dependent samples designs is given below (Dunlap, Cortina, Vaslow, & Burke, 1996):

$$\sigma = \sqrt{[\sum(X - M)^2 / N]}$$

An alternative approach, advocated by Lipsey and Wilson (2001, p. 41–51), is to use the mean scores and standard deviations for Time 1 and Time 2 to calculate the pooled standard deviations and effect sizes.

However, none of the studies included in the present meta-analysis provided original subjects' scores, and most of the dependent-sample studies neglected to report the standard deviations for Time 1 and Time 2. This situation arises frequently in meta-analyses, and one practical solution has been to use a simplified independent samples formula—the *t* score divided by the square root of *N*—as an estimate of effect size for the dependent-sample designs. This approach, which has been followed in studies like Norris and Ortega's (2000) meta-analysis on the effectiveness of L2 instruction, was adopted in the present study. However, it has been shown that computing Cohen's *d* from the *t* score and sample size results is an overestimate of the true magnitude of the effect size (see Dunlap et al., 1996), providing an additional reason why the results for studies with dependent-sample designs should be interpreted with caution in our analysis. We therefore report the results for dependent-sample designs separately from the results for true experimental designs.

In practical terms, a Cohen's *d* of 1.0 means that the treatment group scored one standard deviation higher than the control group (or that there was a gain of one standard deviation from the pretest to the posttest.) Thus, converting all statistical differences to standard deviation units makes it possible to directly compare outcomes across studies.

No absolute standards are used to interpret effect sizes. The most widely accepted rule of thumb, proposed by Cohen (1988), is based on a survey of the typical findings in social science research: Effect sizes of $d < 0.20$ are interpreted as *insignificant*; values of *d* between 0.20 and 0.50 are interpreted as *small effects*; values of *d* between 0.50 and 0.80 are interpreted as *medium effects*; and values of *d* larger than 0.80 are interpreted as *large effects*. However, Norris and Ortega (2006, pp. 33–34) advocated a stricter standard, based in part on their findings in Norris and Ortega (2000), where it seemed that effect sizes around 1.0 were more typical for L2 instructional treatment studies.

Dunlap et al. (1996) showed that effect size estimates based on *correlated designs* (i.e., pretest versus posttest designs) will systematically overestimate the true effect, unless adjustments are applied. Specifically, they found an overestimate by a factor of 2 for studies with a correlation of .75 for the test-retest reliabilities (the typical case; see Dunlap et al. 1996, p. 171). Using this adjusted rule of thumb results in higher required effect sizes for dependent-sample designs: $d < 0.40$ is interpreted as insignificant; *d* between 0.40 and 1.00 is interpreted as

small effects; *d* between 1.00 and 1.60 is interpreted as medium effects; and values of *d* larger than 1.60 are interpreted as large effects.

One major methodological issue for meta-analysis concerns whether it is appropriate to compute multiple effect sizes from a single publication or study. For example, many studies include multiple treatment groups (e.g., that receive different kinds of feedback) where each treatment group is compared to the same control group that received no feedback. Other studies use a single treatment group and a control group, but these two groups are compared with respect to multiple outcome measures. In cases like these, it is statistically possible to compute multiple effect sizes, one for each statistical comparison. But in that case, the effect sizes are not truly independent. Including multiple effect sizes from a study provides greater weight to that particular study, which could become a problem if that study was biased in some way.

At the same time, choosing only a single comparison from a given study fails to represent the overall findings of the study and does not provide the basis for comparisons across different meta-analyses. Thus, we decided to provide a comprehensive coverage of all comparisons reported in these studies, at the risk of including multiple comparisons based on a single group.

Specifically, we used the following approach: First, we computed an effect size for every relevant mean difference reported in these studies. In total, we computed 172 effect sizes from the 25 studies included in our meta-analysis, or on average about seven effect sizes per study. These individual effect sizes are given in Appendix B. We then analyzed the independent variables associated with each effect size, to determine whether they represented distinctions that were relevant for the purpose of our meta-analysis. In cases where two effect sizes were associated with a single configuration of independent variables, we computed an average effect size. Appendix C shows the effect sizes used for our final meta-analysis.

For example, Ashwell (2000) used a pretest-posttest design for three different groups of students. Each group received feedback focused on form and content. The different kinds of feedback were provided in different orders, but those distinctions were not relevant for the purposes of our meta-analysis. Each of the three groups was then evaluated for two outcome measures: one for grammatical accuracy and one for content. Because the distinction between grammar versus content outcomes is relevant for the purposes of our meta-analysis, these individual effect sizes were retained in the final analysis. That is, each group in the Ashwell

study was used for two different effect sizes in the final analysis: one for a grammar outcome measure and one for a content outcome measure.

The analysis of the study by Berg (1999) is relatively uncontroversial: Three different groups received feedback, contrasted with a control group that received no feedback. The groups were compared for a single outcome measure. Because the three groups were independent samples, we retained all three effect sizes in the final analysis.

In contrast, Bitchener, Young, and Cameron (2005) was based on two treatment groups, each compared with a control group for numerous outcome measures. In this case, the outcome measures (e.g., preposition use, tense use, article use) were all specific indicators of the same underlying outcome type: grammatical accuracy. In addition, each group was measured at different points in time. That is, all individual effect sizes for a group are instances of the same configuration of independent variables. As a result, 12 different effect sizes were averaged for each group, so that only two average effect sizes from this study were used in the final meta-analysis.

Appendix C shows the result of this step of the analysis, displaying each of the final effect sizes (or average effect sizes) used in our final meta-analysis. A large number of the original comparisons from these studies were specific measures of the same underlying parameter, and they were averaged for the meta-analysis. Thus, the 172 individual effect sizes that we computed were reduced to 88 effect sizes used in the final meta-analysis. However, those 88 effect sizes take into account every statistical comparison reported in the original studies.

4.3. Computing Mean Effect Sizes and Dispersion Measures

Once the effect sizes were computed for each individual study, it was possible to compute mean effect sizes for the different feedback conditions. For example, it was possible to compare the mean improvement in writing quality (the mean effect size) for students who received error correction compared to students who received global comments. This comparison was accomplished by simply computing the arithmetic mean for all effect sizes of a given type.

However, a simple comparison of mean effect sizes is not in itself very meaningful without also considering the dispersion of effect sizes around that mean. This calculation is required to determine the extent to which effect sizes vary across comparisons of a given type. For this purpose, we computed 95% confidence intervals:

Test of English as a Foreign Language
PO Box 6155
Princeton, NJ 08541-6155
USA

To obtain more information about TOEFL programs and services, use one of the following:

Phone: 1-877-863-3546
(US, US Territories*, and Canada)

1-609-771-7100
(all other locations)

E-mail: toefl@ets.org
Web site: www.ets.org/toefl

*America Samoa, Guam, Puerto Rico, and US Virgin Islands