Available online at www.sciencedirect.com
ScienceDirect
Journal of Second Language Writing 16 (2007) 255–272
JOURNAL OF SECOND LANGUAGE WRITING

The effect of error correction on learners' ability to write accurately

John Truscott

National Tsing Hua University, Taiwan

Abstract

The paper evaluates and synthesizes research on the question of how error correction affects learners' ability to write accurately, combining qualitative analysis of the relevant studies with quantitative meta-analysis of their findings. The conclusions are that, based on existing research: (a) the best estimate is that correction has a small negative effect on learners' ability to write accurately, and (b) we can be 95% confident that if it has any actual benefits, they are very small. This analysis is followed by discussion of factors that have probably biased the findings in favor of correction groups, the implication being that the conclusions of the meta-analysis probably underestimate the failure of correction.

Keywords: Error correction; Grammar correction; Accuracy; Meta-analysis

In the last decade, there has been a great deal of discussion on the value of correction in writing classes, focused primarily on the proper interpretation of relevant research (e.g. Chandler, 2004; Ferris, 1999, 2004; Truscott, 1996, 1999a, 2004). One virtue of this discussion is that it has brought to the field detailed presentations of two fundamentally opposed views, allowing interested parties to compare and make their own informed judgments. More recently, however, the discussion has been largely one-sided, with a wealth of mainstream sources presenting a favorable view of correction and only one brief paper offering an opposing view (Truscott, 2004). Readers could thus be forgiven for thinking that the matter has largely been settled and that the empirical case against correction can now be safely dismissed. Nothing, I will argue, could be further from the truth.

The goal of this paper is thus to show that research evidence points strongly to the ineffectiveness of correction. To this end, I will update the empirical case against the practice, critique recent claims to the effect that research has not really found it unhelpful, and introduce to the discussion a more quantitative dimension, consisting of a small-scale meta-analysis along the lines of Norris and Ortega's (2000) large-scale project on the effects of form-focused instruction

E-mail address: truscott@mx.nthu.edu.tw.

1060-3743/$ – see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.jslw.2007.06.003

(see also Krashen, 2002). I will conclude that correction most likely has small harmful effects on students' ability to write accurately and that we can be reasonably confident that if it does have any genuine benefits, they are so small as to be uninteresting.

1. Some preliminaries

1.1. Meta-analysis and its use

The overall goal of establishing the ineffectiveness of correction entails two sub-goals, each requiring quantification of the research findings. The first is

(1) to find the best estimate of the overall effect of correction on accuracy.

This estimate will be an average based on all the relevant research. Because any estimate is just that, an estimate, the second requirement is

(2) to determine an upper limit on how helpful correction might be.

This limit is probabilistic, as no amount of information could ever produce certainty. The goal is to obtain a number which allows an acceptable degree of confidence that any benefits of correction are no greater than that number, based on existing research. Following standard practice, I will set this confidence level at 95%. The overall goal, then, is a conclusion of this type: ''Based on existing research, the best estimate of the effect of correction on students' ability to write accurately is *X*, and we can be 95% certain that any benefits produced by correction are no greater than *Y*''.

These goals can be achieved through a meta-analysis (see Hedges & Olkin, 1985; Light & Pillemer, 1984; Lipsey & Wilson, 2001; Norris & Ortega, 2000; Rosenthal, 1991, 1994; also Ellis, 2000). The first step is to determine for each study one or more *effect sizes,* which measure how large an effect the independent variable (correction in this case) had on the dependent variable (accuracy of students' writing). I will rely on the measure most widely used, *Cohen's d*, which is the number of standard deviations by which the means of two groups differ. Thus, if an experimental group's mean is one standard deviation above that of a control group, the resulting *d* will be 1.0. Cohen (1992) offered the following rule of thumb for interpreting d: small effect = .20–.50; medium effect = .50–.80; large effect = .80 and up. There is no maximum possible value, and numbers over 1.0 are common. Negative values are interpreted the same way but represent an inverse relation between the variables; if a study looking at the effects of correction on accuracy yielded an effect size of −.25, this would indicate that correction had a small harmful effect. Throughout this review, I will follow this convention: positive effect sizes indicate beneficial effects of correction; negative effect sizes indicate harmful effects.

One virtue of using effect sizes is that they can be averaged to produce an overall estimate of the effects of the independent variable across all the studies and to make comparisons between groups of studies. The resulting *average effect size* is only an estimate of the real average effect size (that which would result if all possible studies were done), so another measure is commonly added to determine the precision of the estimate. This is the *confidence interval*, defined as the smallest interval around the observed average effect size that has a 95% chance of containing the real average effect size; in other words, the likelihood that the actual value lies outside the interval is only 5%. This 5% chance usually includes a 2.5% chance of it being above the interval and a 2.5% chance of it being below. For the present case, though, goal (2) dictates a different logic, as the lower limit is not interesting. The goal is to determine with 95% confidence that the

actual effect size is not above a given number, so there should be a 5% chance of the actual average being *above* the interval.

In the meta-analysis literature, questions have arisen about cases in which a single study yields more than one effect size (see Lipsey & Wilson, 2001; Norris & Ortega, 2000; Rosenthal, 1994). One can either average all the individual effect sizes found or use only one for each study, representing the average of all those calculated for that study. Norris and Ortega adopted a compromise, using multiple measures from a single study only when they were based on different structures. If a study produced one effect size for past tense and another for wh-questions, for example, both would be included. It is not clear which approach is most appropriate here, so I will report the results for all three.

1.2. The selection of studies to be included

Studies to be considered for this review were identified and selected through a general look at published sources, relying primarily on reviews by Ferris (1999, 2003, 2004) and Truscott (1996, 1999a). These reviews were built around the issue of whether correction should be used in L2 writing classes, and this pedagogical focus has determined the studies to be reviewed, both there and here. First, attention is restricted to those studies that looked at correction done in either writing classes or more general classes that included a writing component. All studies reviewed used authentic writing samples for their measures, rather than grammar exercises, because they provide the most clearly valid tests for the value of correction in writing classes. For the same reason, there has been a strong bias in the debate in favor of studies that examine changes in writing ability over a period of months. One-shot treatments, in particular, have generally gotten little respect in this literature and will not be considered here.

The major exception to the avoidance of one-shot experiments has been Ferris's (1999, 2003, 2004) inclusion of what I will call the *revision studies*, which investigated learners' success in revising an essay after receiving feedback of different types on it (Ashwell, 2000; Fathman & Whalley, 1990; Ferris & Roberts, 2001). My own review (Truscott, 1996, 1999a) excluded this type of study, not just because of its short-term nature, but mainly because it offers no measure of changes in students' ability to write accurately, i.e. their learning. Studies of learning look at the difference between a measure of accuracy at one time and a comparable measure done at a later time. A writing task that students do with help from the teacher (the revision) is obviously not comparable to one they do on their own (the original essay), and so a study with this design does not yield any measure of learning, short-term or otherwise, and the revision studies do not address the question in which I am interested. The same logic dictates exclusion of other lines of research on revision, such as Adams (2003) or Paulus (1999).

I have also omitted several sources that Ferris (2003, 2004) and Ferris and Roberts (2001) cited as having found that corrected groups improve, because in those sources I do not find descriptions of studies that could produce such evidence. Ferris (1995a) presented a survey of students' views, as did Ferris (1997), except that she also noted that students showed some success in using corrections to revise a piece of writing. Ferris (1995b) briefly mentioned unpublished work, simply stating that improvement occurred. Other studies are unpublished or otherwise not readily available, again with minimal description offered of their contents.

I separate experiments into two types and analyze them separately. The more important of the two makes controlled comparisons between the writing of students who received correction and that of students who did not. The second type does not include a meaningful comparison, but does offer information on absolute gains made by correction groups. In each case, the meta-analysis

necessarily includes only a subset of the studies that are discussed more qualitatively, because in some cases authors did not provide the information required for effect size calculations.

1.3. The question of error types

Conclusions on this topic are usually stated very broadly, using the term *error correction* and thus including all error types. But distinctions among the various types are interesting. The case against grammar correction (Truscott, 1996, 1999a) was specifically a rejection of *grammar* correction, excluding, for example, spelling errors, which on a priori grounds are among the most correctible error types because they are relatively simple and can be treated as discrete items, rather than integral parts of a complex system (Truscott, 2001).

In this light, consider Lalande's (1982) findings. He included 12 error categories: 11 grammatical types plus orthography. Of the total gains by his experimental group, 83% occurred on the latter. Gains in the grammatical categories were essentially zero. One implication is that errors in orthography might benefit from correction. (An alternative explanation is avoidance, as it is relatively easy to avoid particular words; see below.) Another is that grammar correction has gotten a free ride in some interpretations of Lalande's results. A finding that corrected students improved in orthography but not in grammar should be treated as such.

Unfortunately, the existing research literature rarely allows such distinctions in evaluations of the evidence, so very little material is available for a meta-analysis of different error types. But the literature does suggest, at least weakly, that grammar errors must be distinguished from other types. Fazio (2001) obtained negative effect sizes from a study of two grammatical error types. Sheppard's (1992) negative findings came from verb form errors and an aspect of punctuation that relied on the grammatical notion of clause. Polio, Fleck, & Leder (1998), who found no effect for correction, excluded errors of spelling and capitalization. In contrast, Kepner (1991) looked at surface level errors in general and obtained a more positive effect size; it was small and nonsignificant, but still much greater than that found in the studies that focused on grammar.

The contrast between the findings of Frantzen (1995) and Chandler (2003) is also interesting. Each used two correction groups, one of which devoted far more time than the other to their errors. Frantzen's all-grammar measure found no difference between groups, while Chandler's all-inclusive measure showed a dramatic advantage for the more error-oriented group. If correcting grammar errors has no effect, then a measure that looks specifically at improvement in grammar errors (Frantzen's) should not show any effect for correction. If correcting some non-grammatical errors does have an effect, then a measure that includes these non-grammatical errors (Chandler's) should show an effect for correction. These expectations match the findings, providing additional evidence that correction may have value for some non-grammatical errors but not for errors in grammar. In the absence of more detailed information, though, these observations are only suggestive. There is clearly a need for focused research.

2. The evidence: controlled experiments

The main evidence comes from controlled experiments comparing the effects of correcting with those of not correcting. This distinction should not be confused with that between correcting errors and providing no feedback at all. No one, to my knowledge, recommends the latter policy. Provision of comments on content and clarity, for instance, appears to be universally accepted,

regardless of one's position on error correction. A question of pedagogical interest, therefore, is how classes that use only these other forms of feedback compare to those that use error correction. I will first briefly review the relevant studies and then combine them, with the aim of achieving the goals established above.

2.1. Sheppard (1992)

Sheppard (1992) compared an ESL group that received extensive correction over a 10-week period to one that had identical instruction but received only content-oriented comments, including marginal statements saying when a portion of the writing was difficult to understand. Students in each group had individual conferences with the instructor, in which they talked entirely about their errors (the correction group) or entirely about meaning (the content group). In the final results, the content group had significantly higher scores in marking sentence boundaries, yielding a large effect size (d) of $-.939$. On accuracy of verb forms, the content group obtained nonsignificantly higher scores ($d = -.478$). Thus, correction in this study was not only ineffective but also probably harmful to students' learning, relative to providing feedback only on meaning.

Ferris (2003, 2004) classified this experiment as a study of the differing effects of coded and uncoded feedback, or simply as uncontrolled. The claim, if I understand it correctly, is that Sheppard's (1992) content group received feedback, and so the study does not address the effectiveness of correction. But however one labels this condition, it clearly falls on the no-correction side of the disagreement; i.e., it is very natural for those who reject error correction and bears little resemblance to anything recommended by advocates of correction. Thus, the finding that it produced far superior results is evidence, important evidence, on the effectiveness of correction in writing classes.

Another criticism that has been offered of Sheppard's (1992) study is that the instruction included student–teacher conferences, in which the feedback was discussed (Ferris, 2003, 2004). Therefore, the argument goes, we cannot judge whether the effects resulted from the feedback or the conferences. This argument would seem to apply equally to studies that included rewriting (or error logs or grammar mini-lessons): in such cases we cannot tell whether the effects were produced by the feedback or by the rewriting. But to the best of my knowledge no one ever has, or ever would, put forth such an argument, because rewriting is recognized as a natural part of the feedback process, and so a finding that correction-plus-rewriting is (is not) helpful is clearly a finding that correction is (is not) helpful. The same is true of Sheppard's conferences. Teacher–student discussion of feedback is not a separate treatment, independent of the feedback process; it is a natural part of that process. The fact that Sheppard talked to his students about the feedback he gave them does not in any way limit the meaningfulness of his findings.

2.2. Kepner (1991)

This study randomly assigned intermediate Spanish FL students to an explicit correction group and a message group with no correction. After 12 weeks, no significant difference in errors was found. Ferris (2003, 2004) presented the study as (weak) evidence favoring correction, because the correction group made fewer errors. But the difference between means was only 6.56, an unimpressive contrast given the pooled standard deviation of 16.89. The effect size, .388, falls in the middle of the ''small'' range. In other words, this study found little or no value for correction.

2.3. Semke (1980, 1984)

Semke (1980, 1984) carried out a large 10-week study of the journal writing of third quarter German students at a U.S. university. Students were divided into four groups: direct correction, coded feedback with self-correction by students, comments on content only, and a combination of direct correction and comments on content. She found no significant differences in accuracy between the three correction groups and the comments group. Ferris (2004) concluded that the results were inconclusive because Semke (1984) did not report absolute gains, only the final comparisons, i.e. relative gains. Semke (1980) did report these numbers: the average gain for the comments group was 7.1, compared to an overall average for the correction groups of 6.9. Thus, the study found correction ineffective. The overall effect size for the three correction groups versus the comments group was .107, well below the level at which the effect could even be classified as ''small''. I suggest that these results actually understate the failure of correction. The correction groups all wrote more slowly, presumably indicating that they were paying more attention to accuracy, as they had been trained to do. But they still did not write more accurately than uncorrected students, placing an exclamation mark on the lack of differences and the ineffectiveness of correction that it indicates.

2.4. Polio et al. (1998)

Polio et al. (1998) did a one-semester study in an ESL writing course for graduate and undergraduate students. Their experimental group received correction, grammar reviews, and training in editing their writing, while controls received none of these. Effects were measured by an in-class essay and an in-class revision of that essay, each using two very closely related measures. None produced any significant contrasts. The essays yielded negligible effect sizes of .099 and .104, while those for the revision were −.122 and −.054, for an overall average effect of .007. For the meta-analysis I have used only the latter number, with some misgivings. The revisions and the original essays might be treated as separate measures, but the authors found high correlations between them and expressed doubts as to whether they were measuring distinct constructs.

2.5. Fazio (2001)

Fazio (2001) looked at the effects of correction on accuracy in the journal writing of Grade 5 students in a French-language school over a period of almost 4 months. The study included both native and non-native speakers; I will consider only the latter. One of the three groups received focused correction on two aspects of French grammar. The second received only comments on content, while the third got a combination of these treatments. Classification of this study is confused by the fact that all the students received extensive correction elsewhere in the class. Thus, one might treat all three groups as correction groups and ignore the different treatments they received on their journals. The study would then be classified as an uncontrolled experiment, in which correction had consistently negative absolute effects (see below). I have classified it as a comparison between correction and no correction groups because the journal writing, on which the conditions differed, was quite substantial and the measure used was specifically performance on journals.

All three groups declined in accuracy, with no significant differences among them. The performance of the comments group was somewhat better than that of the correction group ($d = -.378$) and much better than that of the combination group ($d = -.759$). The implication is that the correction harmed students' accuracy, as all groups received correction in the class and

all declined in accuracy, and the group that did not receive it on their journals had clearly the smallest decline on the journals.

Fazio (2001) suggested that these poor results might have resulted from the limited attention students paid to the corrections, as reported by students themselves, rather than from any inherent problems with correction. A problem with this alternative explanation is that students' attention to the corrections appears to be *inversely* related to their performance: the combination group reported paying much more attention than the correction group, but their accuracy scores were considerably lower. These results are consistent with the view that students who more carefully attended to the corrections harmed their learning by doing so. They do not fit well with a claim that corrected students' poor overall performance can be explained by their limited concern with the corrections. Also, students' limited attention is an inherent problem for correction, not a special feature of this study (see Truscott, 1996). Fazio also argued that the failure of correction might be attributed to the nature of the class, but this argument was based entirely on the view that Fathman and Whalley (1990) provided evidence of correction's success in another type of class. I discussed and rejected this view above.

2.6. *Robb, Ross, & Shortreed (1986)*

This study used four groups of EFL students at a Japanese university, each receiving a distinct type of feedback: (a) explicit correction; (b) coded correction; (c) highlighting; and (d) a marginal count of errors in each line. All groups showed significant gains and no significant differences were found among them. I argued previously (Truscott, 1996) that while the experiment was not strictly controlled, it was equivalent to a controlled study because the information presented to group (d) was so limited that it could not have been helpful, so this group can be treated as a no-correction (control) group. The finding that they matched the genuine correction groups (and actually made slightly larger gains than any of them) is then evidence of the ineffectiveness of correction. The authors did not provide the information needed for effect size calculations, but the reported results make the failure of correction clear.

No one, to my knowledge, has explicitly challenged the claim that group (d) was in effect a control group. Pro-correction arguments (e.g. Chandler, 2003) simply note that all groups received feedback and all improved, which can be readily explained by other factors. The experiment lasted from April to January. Given this lengthy period, during most of which the students took a writing class, it would be quite surprising if there were no significant gains, with or without correction.

Experimental support for my original interpretation can be found in a study by Lee (1997), who looked at learners' ability to correct errors implanted in a text presented to them under three different conditions. One was a standard control condition, providing no information about the errors. For another group errors were underlined, corresponding to Robb et al.'s (1986) condition (c). For the third, learners were told which lines contained an error, comparable to Robb, Ross, and Shortreed's group (d). Lee (1997) found underlining quite helpful: learners corrected 50.5% of the errors, versus only 19% for the control group, a significant difference. This result is consistent with findings by Fathman and Whalley (1990) and Ferris and Roberts (2001). In contrast, the ''margin'' group was indistinguishable from the control group, successfully correcting only 22.1% of the errors. The implication is that this condition provided no useful information for identifying and correcting the errors. Returning to Robb et al. (1986), we have confirmation that a group receiving such feedback (d) receives no useful information and so can be treated as a control group. The fact that it performed as well as genuine correction groups is thus evidence that correction was irrelevant to the results.

If one adopts the alternative view, attributing the gains to correction, it appears to be a remarkable coincidence that four radically different types of feedback yielded no differences in outcome. Another implication of such a view is that the marginal tally approach is entirely appropriate for use as the standard form of feedback for an ordinary writing course over a lengthy period of time, a view that does not fit well with standard recommendations. This study offers undisputed evidence that the marginal tally approach should be treated on a par with other, more popular methods, or (more to the point) that they should be treated on a par with it.

2.7. Summary of the controlled experiments

The findings of the controlled experiments are summarized in Table 1, which excludes some measures that are not directly related to writing accuracy, particularly measures of fluency and complexity of writing.

Table 1
Comparison between correction groups and non-correction groups on accuracy

	Groups	*n*	Mean	S.D.	Effect sizes (Cohen's *d*)	Notes
Sheppard (1992)	Correction	13	n.a.	n.a.	−.478	Calculated from *t*-scores
V forms	Content	13	n.a.	n.a.		
Sheppard (1992)	Correction	13	n.a.	n.a.	−.939	
clause boundaries	Content	13	n.a.	n.a.		
Kepner (1991)	Correction	30	37.87	17.90	.388	Means are error counts
	Message	30	44.43	15.89		
Semke (1980, 1984)	Direct correction	27	78.44	n.a.	.107	Calculated from *t*-score comparing comments group to the other three (Semke, 1980, Table X.6); means are accuracy scores
	Coded correction	38	74.12	n.a.		
	Correction + comments	30	75.02	n.a.		
	Comments	46	74.88	n.a.		
Fazio (2001)	Correction	16	.130	.123	−.378	Means are error scores
	Comments	15	.095	.060		
Fazio (2001)	Combination	15	.155	.098	−.759	
	Comments	15	.095	.060		
Polio et al. (1998)	Correction+	34	.311/.249	.132/.127	.007	Essay and revision each used two closely related measures; d is the average for all four measures; means are accuracy scores
essay	No correction	31	.296/.235	.171/.144		
Polio et al. (1998)	Correction+	34	.336/.279	.144/.141		
revision	No correction	31	.355/.287	.169/.156		
Average (confidence limit)						
By measures[a]					−.204 (.094)[d]	
By studies[b]					−.155 (.289)	
By studies/structures[c]					−.247 (.162)	

[a] Includes all the effect sizes in the table and counts Semke's as three, because it represents the average of three comparisons between correction groups and the comments group (nine *d*'s).

[b] Uses the average effect size produced by each study (five *d*'s).

[c] Counts Sheppard's two measures separately and Semke and Fazio each as one, using averages (six *d*'s).

[d] This limit was produced by an unorthodox calculation. Using Semke's average three times, as was done in the calculation of the overall average, would artificially lower the standard deviation, and therefore the confidence limit, which would have been .076 in this case. I compensated as best I could by using Semke's effect size only once in the calculation of the standard deviation.

The average effect size was calculated by the three methods described above. All produced negative means, so the best estimate is that correction has (very small) harmful effects on students' ability to write accurately. The confidence limits are .094, .289, and .162. In other words, one calculation method allows 95% confidence that correction has no better than a small beneficial effect on accuracy, while the other two allow 95% confidence that any beneficial effects are too small to even qualify as small effects. Given these results, the issue appears to be not whether correction is effective, but whether it is merely ineffective or (more likely) mildly harmful.

3. Additional evidence: absolute gains by corrected students

Some studies have not included control groups, instead looking only at absolute gains made by groups receiving correction. The limits of such evidence are clear: in the absence of a control group, one cannot determine whether observed gains resulted from correction or from other factors. Thus, even if corrected students consistently showed significant improvement in their accuracy, this finding in itself would tell us nothing about the value of correction. The way to draw implications from uncontrolled studies is to quantify the gains they find in a way that does allow comparison with general standards and with gains expected in the absence of correction. In this section I will discuss the various studies and present effect sizes for them wherever possible, comparing accuracy before and after the treatment (see Dunlap, Cortina, Vaslow, & Burke, 1996; Light & Pillemer, 1984, p. 56; Lipsey & Wilson, 2001; Norris & Ortega, 2000, p. 446). These will then be combined and evaluated. Consistency with the conclusions drawn from the primary evidence, controlled experiments, will then further support those conclusions.

3.1. The uncontrolled experiments

3.1.1. Hendrickson (1981)

Hendrickson's (1981) study involved a heterogeneous sample of adult learners in an ESL class over a period of 9 weeks. No significant differences were found between the effects of comprehensive correction and correction of global errors only. The author did not present any numbers for absolute gains, so his results cannot be included in the meta-analysis. He was, however, quite negative about the outcome, suggesting that the effect sizes were small.

3.1.2. Lalande (1982)

Lalande (1982) studied intermediate German students at a U.S. university, comparing one group that received coded correction and used error logs to track their errors to another that received explicit correction without logs. The former showed nonsignificant gains, yielding a modest effect size of .288. (And recall that these gains were almost entirely in orthography.) The other correction group had nonsignificant declines in accuracy ($d = -.342$).

3.1.3. Frantzen (1995)

Frantzen (1995) did a 15-week study with intermediate Spanish learners in a university setting in the U.S. She included an uncoded-correction group and a ''grammar group'' that also had extensive grammar reviews and was expected to correct their errors, with additional feedback from the teacher on these corrections. She found no significant differences between the groups on accuracy in their essays, despite the dramatic differences in grammar treatments, and concluded that a content course, without grammar, is sufficient for accuracy in writing, at least in this case.

She used two composite measures of grammar, involving different weighting systems. One showed significant gains in the combined scores of the two groups; the other did not.

Unfortunately, the information presented does not allow any meaningful effect size calculation. Use of the F score given for the measure that produced significance would be misleading, as the effect size representing this study should be the average of the two measures (as was done for Polio et al., 1998), but no F score was given for the other. More importantly, the use of F or t scores from a repeated measures design such as this one leads to an overestimation (Dunlap et al., 1996). Such a calculation for the reported F score yields a d of .681, which could be only a small overestimate or could be two or three times the actual effect size. Therefore, this study cannot be used for the meta-analysis. The same problem occurs with Sheppard's (1992) report of absolute gains (but not with the comparison between his groups, which did not involve repeated measures).

3.1.4. Chandler (2003)

Chandler (2003) carried out two studies in an ESL reading and writing course for music students, each covering most of a semester. The first used one group that rewrote their compositions on the basis of coded corrections and a second that received uncoded correction and were not asked to revise. The former had significant gains ($d = 1.08$) and the latter nonsignificant declines ($d = -.207$). The author presented this study as a controlled experiment, saying the outcome "shows that to increase accuracy in student writing teachers should give error feedback and require students to make corrections" (p. 290). The main point of my response (Truscott, 2004) was that an experiment in which both groups received correction cannot support such a conclusion. Chandler's (2004) counter-response did not make her position on this point clear. She strongly defended her treatment of the no-revision group as a control group and showed no sign of withdrawing or weakening the strong conclusions originally drawn. But she did "accept the argument that the efficacy of error correction for the accuracy of subsequent writing will only be demonstrated by studies containing a control group which receives no correction. . ." (p. 348). In any case, I have classified this study in the way it must be classified: as an uncontrolled experiment. It is, however, worthy of further attention, which I will give below.

Study 2 used four different types of correction with a single group. Each student received all four types, one on each of the first four (of five) assignments, and was required to revise each assignment based on the feedback. The author reported a significant overall average error reduction of 2.1 per 100 words, yielding a modest effect size of .423. The figure of 2.1 represents the difference between the mean on the first assignment ($n = 36$) and that on the fifth ($n = 29$). As the question being asked was how much students improved, the seven whose improvement was not measured are best excluded. This (proper) adjustment would probably reduce the effect size greatly. Chandler's (2003) Table 8 gives the average error reduction for twenty of the students (those who did all the required revision) following each of the four types of correction. Based on those numbers, the overall average error reduction for these students was only .9. Thus, the figure of 2.1 probably owes a great deal to the seven scores that should not have been included in the pretest mean. Lacking sufficient information on the 29 students actually studied, I will use the numbers Chandler reported and simply note that the actual effect size may well be much smaller.

3.1.5. Bitchener, Young, & Cameron (2005)

This study looked at the effects of correction and conferences on the development of three aspects of English grammar: prepositions, simple past tense, and definite articles. It used three groups of adult learners in an ESL setting over a 12-week period. The first had explicit correction,

individual student–teacher conferences, and 20 hours of English instruction per week. The second received explicit correction without conferences and had 10 hours of classes per week. The third had no correction or conferences and received 4 hours of instruction per week. Because of the dramatic differences in amount of instruction, this study could not be included among the controlled experiments. Surprisingly, the authors presented it as a straightforward comparison of the differing effects of correction plus conferences, correction alone, and no correction, without directly addressing the instruction variable. In describing the design, they noted that all three groups received the same amount of instruction in grammar and writing, but of course the enormous differences in classroom exposure cannot be overlooked.

In terms of gains/losses, the no-correction group was slightly better than the correction-only group, despite the latter's huge advantage in hours of instruction, suggesting that the correction may well have been harmful. The performance of the control group was also somewhat better than that of the correction–conference group on past tense and was not dramatically lower on prepositions. It was only on articles that a clear superiority showed for the combination of correction plus conferences plus an additional 16 hours per week of instruction. And this advantage resulted entirely from a dramatic and unexplained reversal in this group's performance on the final task. They declined in accuracy from the first to the second and from the second to the third task but then made a huge improvement from the third to the fourth (of four), yielding a stunning effect size of 1.52 for that final portion of the study. A similar pattern appears with this group on past tense and (to a much lesser extent) on prepositions. With such peculiar numbers, one must ask if the findings for the last assignment for this group were somehow contaminated by an additional, unknown variable (cf. the discussion by Robb et al., 1986, of a similar but less dramatic case in their own findings).

Overall, then, correction groups clearly outperformed the control group on only one of the six comparisons, despite the enormous advantage they enjoyed in hours of instruction, and this one case comes with a large question mark. This summary of the findings is based on the means and standard deviations presented in Table 2 of Bitchener et al. (2005). Following standard practice, I focus on the changes that occurred in the means for each group from the first task (Week 2) to the last (Week 12). The results of this analysis are sometimes difficult to reconcile with the authors' own reports of their findings. Most strikingly, in regard to past tense, their table shows noticeable gains for the control group and tiny losses for the correction–conference group (with somewhat larger losses for the correction-only group). But the authors' summary, based on their own analysis, is that the combination of correction and conferences proved beneficial on this error type.

I have included both of the correction groups in the meta-analysis of absolute gains. Four of the six resulting effect sizes were negative. Of the positive effects, one was small and the other (articles for the correction–conference group) was large.

3.1.6. Ferris (2006)

Ferris (2006) reported the results of a one-semester study, carried out at a U.S. university some years earlier, in which a mixed group of students all received extensive error feedback in their writing class. The original intent was to consistently use coded correction, but the three teachers involved in the study did not follow this plan, instead adjusting their responses to perceived needs of the students. So the feedback is best characterized as ''mixed correction types.'' Fifteen error categories were used and changes in error rates for the five most common types were calculated. Significant improvements were found for the overall rate and for one of the individual types (verb errors). The overall effect size was a ''small'' .336. For the five types individually, three of the

Table 2
Absolute gains in accuracy for correction groups

	Groups	*n*	Mean (S.D.)		Effect sizes (Cohen's *d*)	Notes
			Pre	Post		
Lalande (1982)	Coded	30	28.47 (8.30)	25.23 (14.19)	.288	Means are error rates
	Explicit	30	27.73 (14.43)	32.87 (15.61)	−.342	
Chandler (2003) Study 1	Coded + revision	15	7.8 (3.2)	5.1 (1.8)	1.080	Means are error rates
	Uncoded	16	6.0 (4.1)	6.9 (4.6)	−.207	
Chandler (2003) Study 2	Mixed corr types	36	10.1 (5.5)	8.0 (4.3)	.423	
Fazio (2001)	Correction	16	.101 (.047)	.130 (.123)	−.341	Means are error rates
	Combination	15	.098 (.064)	.155 (.098)	−.704	
Polio et al. (1998) essay	Correction	34	.263/.208 (.14/.127)	.311/.249 (.132/.127)	.409	Means are accuracy scores
Polio et al. (1998) revision	Correction	34	.268/.219 (.127/.122)	.336/.279 (.144/.141)		
Bitchener et al. (2005) prepositions	Conferences	19	82.21 (7.03)	84.79 (8.92)	.324	Means are accuracy scores
	Corr only	17	83.71 (10.56)	75.79 (10.09)	−.767	
Bitchener et al. (2005) past tense	Conferences	19	91.64 (8.66)	91.50 (11.22)	−.014	
	Corr only	17	81.07 (18.49)	77.86 (15.91)	−.187	
Bitchener et al. (2005)	Conferences	19	63.00 (37.22)	83.93 (14.40)	.811	
definite articles	Corr only	17	69.29 (30.60)	61.93 (17.57)	−.306	
Ferris (2006) overall	Mixed corr types	55	52.93 (22.70)	46.25 (17.03)	.339	Means are error rates
Ferris (2006) verb errors	Mixed corr types	55	12.93 (8.95)	9.46 (5.27)	.488	
Ferris (2006) noun errors	Mixed corr types	55	6.20 (4.85)	5.16 (4.46)	.223	
Ferris (2006) article errors	Mixed corr types	55	5.82 (5.46)	6.29 (5.13)	−.089	
Ferris (2006) lexical errors	Mixed corr types	55	13.58 (7.72)	11.78 (8.14)	.227	
Ferris (2006) sentence errors	Mixed corr types	55	13.34 (8.14)	13.39 (8.00)	−.006	
Average (confidence limit)						
By measures[a]					.069 (.259)	
By studies[b]					.148 (.412)	
By studies/structures[c]					.115 (.265)	

[a] Includes all the effect sizes in the table except Ferris's overall value (19 *d*'s).
[b] Uses the average effect size produced by each study, but the overall value for Ferris (seven *d*'s).
[c] Uses one *d* (the average) for each study except Bitchener et al., for which three effect sizes were used representing the average of the two groups for each grammar point, and Ferris, for which all were used except the overall value (13 *d*'s).

effect sizes fell in the ''small'' range and the other two had very small negative values. The average for the five types was a ''negligible'' .169. The other findings of this study indicated that the students were very serious about dealing with the corrections they received (much more so than was found in previous studies) and were quite successful in incorporating them in their subsequent revisions. Thus, these very modest results were obtained under unusually favorable circumstances.

3.2. Absolute gains and their implications

Information on absolute gains by correction groups in uncontrolled studies can be supplemented by that from the controlled experiments described above. One can then determine the typical gains made by correction groups and compare them to gains expected in the absence of correction. The results are shown in Table 2, which of course excludes those studies in which the required information was not provided.

For Fazio (2001), the two groups included are, unambiguously, correction groups, as they received correction on their journals and elsewhere in the class. The performance of the comments group, however, reflects the combination of no correction on journals (the target of the study) and correction elsewhere in the class, so I excluded it. Average effect sizes were calculated in the same ways described in the discussion of controlled experiments.

For all the studies taken together, the averages are extremely small: the best estimate is a negligible effect. And these numbers do not represent the effect of correction, but rather the combined effect of all factors influencing accuracy, including correction. The gains made by Polio et al.'s (1998) correction group, for example, were matched by their comparison group, indicating that correction actually made no contribution. I have included confidence limits, all falling in the small range, but limits calculated from pretest–posttest designs are not trustworthy (Lipsey & Wilson, 2001), so great caution is required.

One could in principle do a general comparison between the averages in Table 2 and those for no-correction groups. But so little information is available for the latter that any such comparison would be extremely unreliable. The averages can also be compared with the average effect size Norris and Ortega (2000) obtained for control groups in their survey, .30. These groups gained not from any treatment but rather from all the other factors that influence absolute gains, such as language experience, maturation, and the benefits of doing a posttest task similar to a pretest they have already done. The fact that this number is greater than the average gains made by corrected groups thus suggests that these extraneous factors are more than sufficient to explain the gains in Table 2, without any appeal to hypothetical benefits of correction. Caution is required, because the research used by Norris and Ortega differed from the correction studies in various ways. It is clear, though, that the gains made by correction groups are very small, even when the various beneficial factors have not been factored out. The results are thus consistent with the negative findings from controlled experiments and lend further support to those findings.

4. Problems in the research: inflating factors

Ferris (2003, 2004) pointed out various problems with the studies, problems that are standard in research on language teaching. But no one has suggested that these flaws introduced any systematic bias against correction groups. So they cannot begin to account for the overall pattern of failure for correction. In addition, two factors probably *have* systematically biased the findings in favor of correction groups, making them look better than they actually are. In this section I will

consider these factors and then take a closer look at Chandler's (2003) Study 1, which raised the average considerably because it produced an effect size far greater than that of any of the other uncontrolled studies, apart from one of the six effect sizes for Bitchener et al. (2005), which I discussed above.

4.1. The setting of the testing

One biasing factor is the reliance on measurements done as part of the class, in the same setting in which the correction was done and with no break between the correction period and the testing. Students are prone to forgetting over time, especially when they no longer have the same teacher, context, and learning tasks reminding them. No one, to my knowledge, has systematically investigated this effect specifically for correction. But in research on grammar instruction, the decline in effects after instruction is well established. Norris and Ortega (2000) found that in 22 studies the difference in average effect size between immediate and delayed posttests was a substantial .22. The delay period in these studies was typically short (a few days to a few weeks) and follow-up testing was normally done in the same context as the instruction, so the actual long-term decline following instruction (correction) might be much larger. Thus, the actual ability of corrected students is probably well below the (already low) levels found in the research. Note also Leki's (1991) observation that students who had avoided certain errors all through a writing course would go back to making them when writing evaluations at the end of the course, apparently because their focus had switched from form to content. I believe a great many teachers could offer similar observations from their experience.

4.2. Avoidance

A second likely source of bias in the findings is avoidance. Corrected students tend to shorten and simplify their writing (Kepner, 1991; Semke, 1980, 1984; Sheppard, 1992), apparently to avoid situations in which they might make errors. And this finding should be expected. The existence of the phenomenon was established long ago, in the context of error analysis (e.g. Kleinmann, 1977; Perkins & Larsen Freeman, 1975; Schachter, 1974). The observation is that learners who find a construction difficult tend to avoid it, using it only when especially confident that they can get it right, or when they have no choice. For grammar, the possibility of paraphrase means they very often do have a choice (Schachter, 1974).

This observation has clear relevance to correction, the immediate goal of which is to make learners aware of their errors. Learners are often confused about the corrections they receive (e.g. Lee, 2004; see also Truscott, 1996). When they do understand, this does not mean they have gained mastery of the corrected form, especially of how to apply it to other contexts (see Odlin, 1994; Truscott, 1996). So in many cases, a natural reaction to a correction is ''I have a problem, but I'm not really clear about it''. This awareness creates a clear motivation for avoiding the type of construction corrected. The frequent uncertainty about the exact nature of the problem should also produce the broader type of avoidance found in the research, as learners cannot be sure when they are stepping into dangerous territory.

The implication is that corrected students hide their weaknesses. So when their scores rise on overall accuracy, this apparent improvement might simply mean they have learned to avoid using things they might get wrong. Improved scores on a particular form might mean they now use that form only when most confident about getting it right. Uncorrected students should also show avoidance, but to a lesser extent, as they are not repeatedly pushed to focus on their errors. The

phenomenon of avoidance thus suggests that research findings often overestimate corrected students' ability, both absolutely and in relation to uncorrected students. Unfortunately, it is difficult to judge the strength of the effect or the specific cases in which it is likely to be a serious problem. This is clearly a worthwhile area for further research.

Signs of avoidance are not hard to find in correction studies. Sheppard (1992) found clear evidence, in the form of significant drops in complexity. Frantzen (1995, p. 344) noted that "Several students managed not to generate any contexts for the subjunctive on one or both of the essays," even though they were required to write at least eight paragraphs per essay and directions pushed them to use subjunctive (and other forms). For this measure, she ended up using only 27 of the 44 subjects, apparently for this reason. This finding screams "avoidance." It also fits well with the lack of progress made in fluency over the 15-week writing class. An old observation, which to my knowledge has never been challenged, is that the length of writing students do (when given a choice) is a function of their proficiency and that increases during a writing class are normal (see Hendrickson, 1981, and references cited there). Thus, small or non-existent increases, as in Frantzen's study, suggest substantial avoidance.

Avoidance is also likely in Chandler's (2003) study (see Truscott, 2004). Chandler (2004) argued that it can be ruled out if one counts *all* errors and that avoidance of an error type *is* an increase in accuracy. Such claims misunderstand the phenomenon. It is not errors as such that are avoided but rather the situations in which one might make them. A learner afraid of forming relative clauses incorrectly can avoid them by means of paraphrasing; a learner who has trouble spelling certain words can use other words in their place. Writers who come to use such strategies are no doubt learning to avoid some error types, but this does not mean they are becoming better writers, in any sense. Reliance on avoidance strategies is in all likelihood a major barrier to such progress. Thus, declines in error rates by Chandler's correction–revision group may well reflect, to an unknown extent, a counterproductive risk-avoidance strategy rather than any progress in their writing ability.

4.3. *Chandler's (2003) Study 1*

The effect size of 1.08 produced by the correction–revision group of Chandler's (2003) Study 1 greatly influenced the averages in Table 2. So one must ask how much of this effect can be attributed to correction. The main answer is that we have no way of knowing; because of the nature of the study, only conjecture is possible (Truscott, 2004).

Chandler's (2003, 2004) conjecture was that the gains were produced by the combination of correction and revision. The failure of this combination to produce success in other studies (Polio et al., 1998; Robb et al., 1986; Semke, 1984; Sheppard, 1992) should raise initial doubts about this account. The general need for caution in attributing observed gains to correction is shown by Sheppard's (1992) finding of large gains by a correction group accompanied by larger gains by a comparison group. I already described one alternative factor, avoidance, that may explain much of Chandler's outcome, along with a complementary possibility—that the treatment produced genuine improvements but only in some non-grammatical error types. Here I will focus on a third factor, students' experience with English during the period of the study.

This experience included very extensive writing practice, all on one topic, plus extensive English input in the class and a virtual guarantee of much more outside, given the university ESL setting. Chandler's (2004) response to these points did not address exposure within the class or the point that (by her assessment) the writing practice should have greatly helped all the students. For outside exposure, the claim was that students "were not necessarily receiving considerable

exposure to academic English'' (p. 346). Beyond the overt weakness of the claim (''not necessarily''), the restriction to academic English removes most of its force. Non-academic writing also requires paragraph indentation, capitalization, subject–verb agreement, plurals, correct spelling, and indeed all or nearly all the other categories that Chandler studied, many of which are comparable in spoken English as well. Outside exposure remains a very likely factor in the results, as does the extensive experience obtained in the class.

This discussion strongly suggests that the no-revision group was harmed by the correction. These students should have benefited greatly from experience with English, but instead their accuracy scores declined. If correction had a harmful effect, as it did in other cases, this finding is understandable. If it was helpful, or even neutral, there is no apparent explanation. The gap between this group and the correction–revision group could have been produced by several factors. The latter had much greater motivation for avoidance, as they were required to pay much more attention to their errors and to do additional work for each error they committed. Another possible contributor is the rewriting itself, which could have benefited the one group that did it, independent of the correction they received. It is also quite possible that the correction–revision group did gain from the treatment but only in certain non-grammatical error types.

To summarize, the source of the apparent gains made by Chandler's (2003) correction–revision group is a matter of conjecture. An explanation in terms of avoidance, language experience, and rewriting in itself (and possibly very narrow genuine gains from correction plus revision) is at least as plausible as one that hypothesizes broad beneficial effects of correction–revision. Therefore, the meaningfulness of the exceptional effect size found for this group is very much open to doubt.

5. Conclusion

The primary conclusion, based on the controlled experiments, is (a) the best estimate is that correction has a small harmful effect on students' ability to write accurately, and (b) we can be 95% confident that if it actually has any benefits, they are very small. This conclusion receives support from the finding that absolute gains made by corrected students are quite limited, even when beneficial extraneous factors are not controlled. And these negative results probably overestimate the success of corrected groups, especially in regard to grammar errors.

These conclusions must be contrasted with those reported by the authors of another meta-analysis of research on the effects of correction, Russell and Spada (2006). After synthesizing a number of studies, they reported that correction is quite effective, stating that their findings were inconsistent with my conclusion that research has found it ineffective (Truscott, 1996, 1999b). But in fact their findings are entirely consistent with that conclusion. They obtained a different answer than I did because they asked a different question.

My question was (and is) how correction affects learners' ability to use the language in realistic ways—in writing or speaking for communicative purposes. Answering this question means looking specifically at studies that measured changes in this ability following a period of correction. Those that looked only at learners' performance on artificial grammar tests are necessarily excluded because they do not address the question, as are revision studies, for the same reason. In contrast, Russell and Spada's (2006) sample was dominated by studies of these two types. Thus, the question they actually investigated was whether correction has effects of any kind, without regard to what type of learning was being measured in a given study or even whether the study included a measure of learning. It would be remarkable if a meta-analysis of this sort did not yield large positive effect sizes; no one, to my knowledge, has ever doubted that

correction can help students prepare for grammar tests or revise the piece of writing on which they received the corrections. To my thinking, the findings reported by Russell and Spada are simply not interesting, because the question that they address is not interesting. It should also be noted that their meta-analysis did not include some studies that have important implications (all negative) for the answers both to my question and to theirs (Polio et al., 1998; Semke, 1980, 1984; Sheppard, 1992) and that they mistakenly reported Fazio's (2001) negative effect for correction as a positive effect.

Finally, evidence regarding the effectiveness of correction is clearly relevant to teaching, particularly to decisions by teachers to correct or not to correct. But pedagogical issues raise additional questions, which cannot be addressed here, so I will not offer any advice to teachers. Instead, the conclusion is simply that research has found correction to be a clear and dramatic failure. The performance of corrected groups is in fact so poor that the question ''How effective is correction?'' should perhaps be replaced by ''How harmful is correction?''

Acknowledgements

I wish to thank Stephen Krashen and Viphavee Vongpumivitch for some helpful discussion.

References

Adams, R. (2003). L2 output, reformulation and noticing: Implications for IL development. *Language Teaching Research, 7*, 347–376.

Ashwell, T. (2000). Patterns of teacher response to student writing in a multiple-draft composition classroom: Is content feedback followed by form feedback the best method? *Journal of Second Language Writing, 9*, 227–257.

Bitchener, J., Young, S., & Cameron, D. (2005). The effect of different types of feedback on ESL student writing. *Journal of Second Language Writing, 14*, 191–205.

Chandler, J. (2003). The efficacy of various kinds of error feedback for improvement in the accuracy and fluency of L2 student writing. *Journal of Second Language Writing, 12*, 267–296.

Chandler, J. (2004). A response to Truscott. *Journal of Second Language Writing, 13*, 345–348.

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

Dunlap, W. P., Cortina, J. M., Vaslow, J. B., & Burke, M. J. (1996). Meta-analysis of experiments with matched groups or repeated measures designs. *Psychological Methods, 1*, 170–177.

Ellis, N. (2000). Editor's statement. *Language Learning, 50*(3), xi–xiii.

Fathman, A. K., & Whalley, E. (1990). Teacher response to student writing: Focus on form versus content. In B. Kroll (Ed.), *Second language writing: Research insights for the classroom* (pp. 178–190). Cambridge: Cambridge University Press.

Fazio, L. L. (2001). The effect of corrections and commentaries on the journal writing accuracy of minority- and majority-language students. *Journal of Second Language Writing, 10*, 235–249.

Ferris, D. R. (1995a). Student reactions to teacher response in multiple-draft composition classrooms. *TESOL Quarterly, 29*, 33–53.

Ferris, D. R. (1995b). Teaching students to self-edit. *TESOL Journal, 4*(4), 18–22.

Ferris, D. R. (1997). The influence of teacher commentary on student revisions. *TESOL Quarterly, 31*, 315–339.

Ferris, D. (1999). The case for grammar correction in L2 writing classes: A response to Truscott (1996). *Journal of Second Language Writing, 8*, 1–11.

Ferris, D. R. (2003). *Response to student writing: Implications for second language students*. Mahwah, NJ: Erlbaum.

Ferris, D. (2004). The *''grammar correction'' debate in L2 writing*: Where are we, and where do we go from here? (and what do we do in the meantime?). *Journal of Second Language Writing, 13*, 49–62.

Ferris, D. (2006). Does error feedback help student writers? New evidence on the short- and long-term effects of written error correction. In K. Hyland & F. Hyland (Eds.), *Feedback in second language writing: Contexts and issues* (pp. 81–104). Cambridge: Cambridge University Press.

Ferris, D., & Roberts, B. (2001). Error feedback in L2 writing classes: How explicit does it need to be? *Journal of Second Language Writing, 10*, 161–184.

Frantzen, D. (1995). The effects of grammar supplementation on written accuracy in an intermediate Spanish content course. *Modern Language Journal, 79*, 329–344.

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

Hendrickson, J. M. (1981). *Error analysis and error correction in language teaching*. Singapore: SEAMEO Regional Language Centre.

Kepner, C. G. (1991). An experiment in the relationship of types of written feedback to the development of second-language writing skills. *Modern Language Journal, 75*, 305–313.

Kleinmann, H. H. (1977). Avoidance behavior in adult second language acquisition. *Language Learning, 27*, 93–107.

Krashen, S. D. (2002). The comprehension hypothesis and its rivals. *Selected papers from the eleventh international symposium on English teaching/fourth Pan-Asian conference* (pp. 395–404).

Lalande, J. F., II (1982). Reducing composition errors: An experiment. *Modern Language Journal, 66*, 140–149.

Lee, I. (1997). ESL learners' performance in error correction in writing: Some implications for teaching. *System, 25*, 465–477.

Lee, I. (2004). Error correction in L2 secondary writing classrooms: The case of Hong Kong. *Journal of Second Language Writing, 13*, 285–312.

Leki, I. (1991). The preferences of ESL students for error correction in college-level writing classes. *Foreign Language Annals, 24*, 203–218.

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

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.

Norris, J. M., & Ortega, L. (2000). Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis. *Language Learning, 50*, 417–528.

Odlin, T. (1994). Conclusion. In T. Odlin (Ed.), *Perspectives on pedagogical grammar* (pp. 314–317). Cambridge: Cambridge University Press.

Paulus, T. M. (1999). The effect of peer and teacher feedback on student writing. *Journal of Second Language Writing, 8*, 265–289.

Perkins, K., & Larsen Freeman, D. (1975). The effect of formal language instruction on the order of morpheme acquisition. *Language Learning, 25*, 237–243.

Polio, C., Fleck, C., & Leder, N. (1998). "If I only had more time:" ESL learners' changes in linguistic accuracy on essay revisions. *Journal of Second Language Writing, 7*, 43–68.

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

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

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

Russell, J., & Spada, N. (2006). The effectiveness of corrective feedback for the acquisition of L2 grammar: A meta-analysis of the research. In J. M. Norris & L. Ortega (Eds.), *Synthesizing research on language learning and teaching* (pp. 133–164). Amsterdam: Benjamins.

Schachter, J. (1974). An error in error analysis. *Language Learning, 24*, 205–214.

Semke, H.M. (1980). The comparative effects of four methods of treating free-writing assignments on the second language skills and attitudes of students in college level first year German. Unpublished Ph.D. dissertation. University of Minnesota.

Semke, H. D. (1984). Effects of the red pen. *Foreign Language Annals, 17*, 195–202.

Sheppard, K. (1992). Two feedback types: Do they make a difference? *RELC Journal, 23*, 103–110.

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

Truscott, J. (1999a). The case for "The case against grammar correction in L2 writing classes": A response to Ferris. *Journal of Second Language Writing, 8*, 111–122.

Truscott, J. (1999b). What's wrong with oral grammar correction. *Canadian Modern Language Review, 55*, 437–456.

Truscott, J. (2001). Selecting errors for selective error correction. *Concentric: Studies in English Literature and Linguistics, 27*, 225–240.

Truscott, J. (2004). Evidence and conjecture on the effects of correction: A response to Chandler. *Journal of Second Language Writing, 13*, 337–343.