Articles

The Effects of Written Corrective Feedback on Japanese EFL Students' Writing

Darby McGrath
Ritsumeikan Asia Pacific University

This study investigated the effect of written corrective feedback (WCF) on written accuracy. It focused on accuracy in 4 pieces of writing over 2 months in a group of 30 students at a Japanese university. No significant effect was found either for time or feedback type on accuracy. This is contrary to the findings of a majority of recent research. Possible factors in this discrepancy are considered. This article discusses the possibility that WCF is ineffective, or at least less effective, in EFL contexts. It is also speculated that the experimental approach taken in this study may have played a role in its null findings.

　本研究は英作文の正確さについて、英作文に対する様々な訂正法（written corrective feedback以下WCF）の効果について調査したものである。この調査は日本の大学に於いて、三十名の学生を対象に二ヶ月間に渡り、四つの英作文を正確さに焦点を当てながら異なったフィードバックを行ったものである。正確さという点に於いて、回を追っても英作文に対する様々な訂正法の違いに関わらず、有意な効果は見受けられなかった。この結果は昨今における大多数の研究結果とは結論を異にする。このような結果が出た理由を推察してみると次のことが挙げられる。まず、この研究で論じているのはEFLとして学ぶ状況を鑑みた時、WCFは効果がない、もしくは、あまり効果がない恐れがあるということである。また、リサーチデザインによる影響も考慮すべきであろう。

Keywords: EFL; writing; written corrective feedback

T here is an intuitive appeal to written corrective feedback (WCF) in second-language writing instruction. That teachers should offer error correction to help students to improve their writing seems self-evident. In fact, prior to the publication of Truscott's (1996) paper calling for the abandonment of WCF, its use was uncontroversial. Since then, there has been debate. At issue is the question of whether WCF has any long-term effect on students' written accuracy or language acquisition. Although much of the recent research points to some efficacy for WCF (see, for example, Bitchener, 2008; Bitchener & Knoch, 2008, 2009, 2010a, 2010b; Bitchener et al., 2005; Ellis et al., 2008; Van Beuningen et al., 2012), the question remains disputed. Opponents (Truscott, 1999, 2004, 2007, 2009, 2010; Truscott & Hsu, 2008) hold that WCF has no benefits, and may even be detrimental.

The efficacy of WCF is important. Teachers and students devote time to providing and attending to WCF. If it is ineffective, this time is wasted. It is in the interests of all stakeholders that we move closer to understanding the extent to which WCF facilitates learning.

One aim of this study was to contribute to our understanding of the value of WCF as a vehicle for language acquisition in EFL contexts, specifically Japan. Research has pointed broadly to a positive effect for WCF acquisition in ESL contexts, but EFL-based studies are rare and have been less positive. This study also attempted to fill three smaller niches. First, because it was not classroom-based it is one of the few studies for which language input is not a confounding variable. The fact that the participants in this study did not receive language instruction for its duration may be unique in this area of research. Second, many recent studies have been highly focused, inasmuch as they have measured feedback on one language point. Although this focus is interesting from a theoretical perspective, it is arguably of limited use in the classroom because it bears little relationship to what many teachers do or to what their students expect.

Finally, this study aimed to contribute to the debate over two different types of WCF: direct WCF, in which the learner is given the correction, and indirect coded WCF, in which the error type is indicated for the learner, but the correction is not provided. Recent literature has tended to argue for the direct approach (e.g., Bitchener & Knoch, 2010a). However, it is not clear that this research has been conducted in conditions in which we might expect indirect feedback to be effective. This study was designed to provide such conditions to give a better indication of the relative merits of direct and indirect WCF.

The study therefore sought to answer the following research questions:

RQ1. Do Japanese EFL students who correct written errors by way of comprehensive, indirect or comprehensive, direct WCF make fewer grammatical errors over time?

RQ2. Do Japanese EFL students who correct written errors by way of comprehensive, indirect or comprehensive, direct WCF make fewer grammatical errors over time than students who receive no WCF?

RQ3. Do Japanese EFL students who correct written errors by way of comprehensive, indirect WCF make fewer grammatical errors over time than students who correct written errors by way of comprehensive, direct WCF?

What Does the Literature Currently Suggest?

Is Written Corrective Feedback Effective?

Early research into WCF tended not to support its use (see Cohen & Robbins, 1976; Kepner, 1991; Robb, et al., 1986; Semke, 1984; Sheppard, 1992). Truscott (1996) was justified in suggesting that there was, at the time, little evidence for its efficacy in language learning. However, there was also a lack of evidence to suggest that WCF is ineffective. The fact is that the research had, until then, not been of a sufficiently high standard to inform an approach to WCF either way. Since 1996, the evidence for the efficacy of WCF has been more compelling, primarily because the use of control groups has become standard, where it was not earlier. Controlled studies have tended to reflect more positively on the use of WCF.

The ten years after Truscott's (1996) call for the abandonment of WCF were characterised by a series of studies that found a positive effect for WCF on written accuracy. Bitchener et al. (2005), Chandler (2003), and Ferris (2006) all reported significant, if modest, effects for WCF. This sequence was broken by Truscott himself. Truscott and Hsu (2008) conducted a rare study published in the last twenty years to find WCF entirely ineffective for learning.

Bitchener (2008) looked at 75 students at two language schools in New Zealand. He found an effect for WCF over the control group that was still present in the delayed posttest 2 months later. Bitchener's study is typical of a recent preference for focused research that investigates the effects of WCF on a single language point. Bitchener and Knoch (2008, 2009, 2010a, 2010b), Ellis et al. (2008), Sheen (2007), and Shintani and Ellis (2013) have

all published studies with the same focus. All of these studies were limited to the effects of WCF on article use.

Two studies by Van Beuningen et al. (2008, 2012) have diverged from this trend towards focused research. They looked at the effects of WCF on written accuracy for high school students of Dutch. These studies differed from other recent studies in that they looked at comprehensive feedback rather than the focused approach. They also showed an effect for WCF on accuracy.

More recently, Karim and Nassaji (2018) looked at the effects of comprehensive WCF on accuracy on the writing of 53 intermediate-level ESL learners. They found that treatment groups outperformed the control group in revision, but there was no significant effect for new writing.

In the context of the current study it is worth noting how few of the studies discussed above have taken place in an EFL setting. Of the three EFL studies mentioned, two were null (Robb et al., 1986; Truscott & Hsu, 2008) and the other was positive but looked only at the effects of WCF on article use (Ellis et al., 2008). Although findings regarding the effectiveness of WCF in ESL settings are generally positive, if limited, EFL-based research has been less so.

Which Forms of WCF Are Most Likely to Be Effective?

A distinction is made in this study between direct WCF and indirect WCF. With the former, the teacher corrects the error for the student, while with the latter the error types are identified by the teacher, but the student corrects them. Early studies that compared the two forms of feedback found no significant difference (Lalande, 1982; Robb et al., 1986; Semke, 1984). Since then, results have varied. Ferris and Hely (2001, as cited in Bitchener et al., 2005 and Bitchener & Ferris, 2012) found direct WCF was more effective in improving accuracy in revisions, but indirect WCF was more effective for learning. In contrast, Bitchener and Knoch (2010a) found that direct WCF was more effective for learning. Chandler (2003) found that although both types of WCF were effective, direct WCF outperformed indirect. Van Beuningen et al. (2008) found that both direct and indirect WCF were effective, but the effects of direct WCF lasted longer. This contrasts with their later study (Van Beuningen et al., 2012), which showed an interaction between feedback type and error type. They found that direct WCF was better for improving grammatical accuracy, whereas indirect WCF was more effective at treating non-grammatical errors. Finally, Karim and Nassaji (2018) found no significant effect for either form of WCF on learning.

Clearly, this is a confused area. Holistically, the research does not support the position that direct WCF is preferable, but the notion has taken hold. Since 2005 most research has focused on direct WCF only (see Bitchener, 2008; Bitchener & Knoch, 2008, 2009, 2010b; Bitchener et al., 2005; Ellis et al., 2008; Sheen, 2007). There are reasons other than the lack of strong evidence to suggest that this may be premature. A second strand of research has focused on learners' interactions with WCF through think-aloud protocols (Qi & Lapkin, 2001), stimulated recall (Adams, 2003), and reflective discussions (Storch & Wigglesworth, 2010). These studies have provided insights into the ways in which learners engage with feedback and have highlighted the importance of noticing to its efficacy. Qi and Lapkin (2001) found in their case studies of ESL students that it was not just noticing, but quality of noticing that determined uptake. To them, this is determined by the extent to which it is "substantive" rather than "perfunctory" (p. 291).

Similarly, Adams (2003), in a study of 56 ESL students, found that the accuracy of those who noticed and acted upon reformulations improved in a written posttest over those who received direct correction. She attributed this to the passivity with which students engage with direct feedback. Although Adams looked at reformulations rather than the coded WCF used in the current study, implicit in her conclusion is that we should expect a form of feedback that asks more engagement from students, such as indirect coded feedback, to be more effective. Further support for this comes from Storch and Wigglesworth (2010), who investigated the extent to which direct and indirect WCF were noticed by 40 ESL students. They found not only that learners engaged more with indirect feedback, but that "retention seemed to relate to the level of engagement" (p. 327).

These qualitative studies have suggested that we should expect indirect WCF to be more effective for learning than direct WCF, which poses the question of why the results have been inconclusive in the quantitative studies discussed earlier. One possibility is that studies that report a greater effect for direct WCF have not provided a setting conducive to the efficacy of the indirect approach. It seems likely that for indirect WCF to be effective, learners require time to engage with it (Polio et al., 1998; Qi & Lapkin, 2001). Indirect coded WCF therefore requires a setting in which students are given time to reflect on their errors, consider how to repair those errors, and make the necessary changes to their writing. This will engage learners in the "reflective learning processes" (Ferris, 2006, p. 83) that are the appeal of indirect WCF.

Returning to the studies that have directly compared indirect and direct WCF, this possibility receives some support. Bitchener and Knoch (2010a) found that direct WCF was more beneficial than indirect. However, the participants in this study did not revise their writing; they were "given a few minutes to consider the feedback" (p. 213). If we accept that indirect feedback requires time and revision to effect learning, then we would expect indirect WCF to fare poorly in this study, as it did. Moreover, if this approach is compared to studies in which learners revised their writing based on indirect feedback, we see different results. Ferris and Hely (2001, as cited in Bitchener et al., 2005) found that indirect WCF was more beneficial to learning, but Van Beuningen et al. (2012) found that indirect WCF was more beneficial for lexical errors. Both of these studies required the participants to make revisions based on WCF.

This perspective suggests that recent research has prematurely discounted the possibility that indirect WCF is effective. There are good intuitive and empirical reasons to believe that, given the right conditions, indirect WCF might be preferred. This study aimed to create those conditions to test this possibility.

How Focused Should Feedback Be?

The distinction here is between focused WCF (in which one or few error types are treated) and comprehensive WCF (in which many or all error types are treated). Recent research has tended to use focused treatments. Since 2007, at least four studies have focused on the treatment of articles in student writing (Bitchener & Knoch, 2009, 2010a; Ellis et al., 2008; Sheen, 2007; Shintani & Ellis, 2013). This is due largely to a perception that comprehensive WCF is likely to overwhelm students, particularly those at lower levels (Bitchener, 2012; Ellis et al., 2008).

However, as with the preference for direct WCF over indirect, this conclusion may be premature. First, it is worth noting that the belief that comprehensive feedback is overwhelming for students is based largely on intuition. Although it seems likely that comprehensive feedback would have this effect, for writing the research is inconclusive. At least four studies have directly compared the two approaches. Ellis et al. (2008) found that focused and comprehensive treatments outperformed controls equally. Sheen, Wright and Moldowa (2009) found that both the focused treatment group and a control group outperformed the comprehensive WCF group. Finally, Frear and Chiu (2015) found that the focused and the comprehensive treatment groups outperformed the control group equally.

Another reason to question the preference for focused WCF is similar to an issue already raised with regard to indirect feedback. We would expect comprehensive WCF to be more time-intensive for learners to attend to than focused WCF. However, we find that this is not necessarily accounted for in studies that found the focused approach to be more effective. Sheen et al. (2009, p. 562), for example, had students "look over their errors…for a few minutes." It is questionable whether this will be sufficient for learners to assimilate comprehensive feedback. Van Beuningen et al. (2012), on the other hand, had students revise their writing based on the feedback they received, and gave them the time to do so. It is no surprise that the authors found comprehensive WCF effective for learning where Sheen et al. did not.

Focusing on comprehensive feedback also increases the ecological validity of the research. Although the studies concentrating on focused WCF show an effect on language learning, what they represent is arguably so far removed from classroom reality that they are of little practical use to teachers and students. For many teachers, the idea that WCF is only effective if it focuses on one language point would be functionally equivalent to it being ineffective. The suggestion of Ellis et al. (2008) that teachers treat a different language point in each paper they set is no real solution in contexts in which students only write one or two papers a semester. The decision was therefore taken in the present study to focus on comprehensive feedback because it closely approximates the WCF teachers tend to give.

Method

Context and Participants

The study was conducted at a private Japanese university. The university's English Language Program (ELP) is divided into four skills-focused courses, including writing. Writing classes employ indirect coded WCF, using the same symbols as were used in this study to code students' errors (see Appendix A). The programme approaches writing as a process, whereby students engage in an error-correction cycle of WCF and editing. WCF aside, there is little form-focused instruction within the ELP. It focuses instead on building skills through contextualised, and often ungraded, language input.

The participants were 30 undergraduate students. All participants were native Japanese speakers between 20 and 21 years old. Twenty-one were female, and nine were male. All participants had recently completed the ELP and so were familiar with the coded symbols used for indirect marking and with the manner of attending to feedback that the study asked of them. All

had studied English for between 6 and 15 years, although most had studied for 8 to 10 years, starting in elementary school. State English education in Japan tends to be highly form-focused, in contrast to the approach of the ELP. The participants had therefore been educated within two fundamentally different pedagogical frameworks.

The participants were paid volunteers, all of whom provided written informed consent. This approach was taken for two reasons. It allowed for a control group without the ethical concerns this would engender in a classroom study. Secondly, because the participants had already completed the ELP it was possible to ensure that none received any EFL instruction for the duration of the study. This obviated a common confounding variable for research of this nature.

Because of two late withdrawals, the indirect feedback group was the largest (*n* = 11), followed by the direct feedback group (*n* = 10) and the control group (*n* = 9). Based on their results in a recent Test of English as a Foreign Language (TOEFL), the participants were divided into five proficiency groups. The participants within these groups were then distributed randomly among the treatment groups to balance language proficiency. A Kruskal-Wallis rank sum test applied to the TOEFL scores showed no significant differences between the proficiency levels of the three groups, $H(2) = 0.03$, $p = .99$. The same test was applied to the pretest error ratios, and this also suggested no significant disparity, $H(2) = 0.69$, $p = .71$.

Design

This study can be characterised as experimental. Participants were each assigned to one of two treatment groups or a control group. Structured random assignment was used to create the groups, as outlined in *Context and Participants*, above. Recruitment was limited to participants who were not currently receiving formal language instruction.

Table 1 shows how the study was organised. Data were collected on four occasions over two months. In sessions 2 and 3, the treatment was based on asking participants to revise their previous essay based on the feedback they had been given.

Table 1

Task Type by Session

Session	Treatment	Essay writing task
1 (pretest)		✓
2 (practice essay)	✓	✓
3 (posttest)	✓	✓
4 (delayed posttest)		✓

Treatments and Writing Tasks

The indirect WCF group received feedback based on a thirteen-item list of codes that designate the type of error (see Appendix A). The approach to WCF in this study was therefore comprehensive. Table 2 summarises the treatments.

Table 2

Summary of Treatments by Group

Group	WCF received	Revision task
1. Control (*n* = 9)	None	Yes (self-correction)
2. Indirect WCF (*n* = 11)	Error and error-type identified by code	Yes
3. Direct WCF (*n* = 10)	Error identified and corrected	Yes

Note. Appendix A shows error codes. Appendices B and C show examples of indirect WCF and direct WCF, respectively

The tendency in recent research into WCF has been to use narratives or picture descriptions as task types (see, for example, Bitchener, 2008; Bitchener & Knoch, 2008, 2009, 2010a; Karim & Nassaji, 2018; Sachs & Polio, 2007; Sheen, 2007; Sheen et al., 2009; Shintani & Ellis, 2013; Truscott & Hsu, 2008). Another approach has had students write essays or similar texts (Ferris et al., 2013; Van Beuningen et al., 2012). The latter method was used in this study because it elicits larger samples of language and was a task with which the participants were familiar. The four tasks were descriptive essays with the same thematic focus (see Table 3). The participants had encountered a similar theme in the ELP.

Table 3
Writing Tasks

Session	Task
Task one (pretest)	What are the features of a good parent?
Task two (practice essay)	What are the features of a good student?
Task three (posttest)	What are the features of a good friend?
Task four (delayed posttest)	What are the features of a good teacher?

Data Collection

The participants completed four pieces of writing: a pretest, a practice essay, a posttest, and a delayed posttest. They were asked to revise the first two pieces of writing, but not the posttests. The first three sessions took place over the course of 1 month and the delayed posttest was 1 month later.

Session 1 began with a preamble and administration. The participants then saw the essay question and had 7 minutes to generate content ideas in groups. Finally, they were given 30 minutes to type their essay. The first written task served as a baseline measure of accuracy and also was used to provide WCF for the first treatment in the following session. This approach has become common in research of this type (for example, Bitchener, 2008).

In Sessions 2 and 3, essays from the previous session were returned and the participants were given 15 minutes to revise them based on their WCF. The control group self-corrected, as they received no WCF. The remainder of the second and third sessions was spent writing Tasks 2 (practice essay) and 3 (posttest). The approach here was identical to Session 1. Session 4 served as the delayed posttest. The approach to essay preparation and writing in session four was identical to the preceding three sessions, but no revision was required.

Data Processing Procedures and Reliability Measures

All essays were rated by the author and the procedure was blinded. Accuracy was measured by means of error ratios, an approach that has been employed by Truscott and Hsu (2008), Chandler (2003), and Van Beuningen et al. (2012). More specifically, given the possibility that some participants would not reach the target of 250 words per task, Van Beuningen et al.'s technique was used because it is appropriate for texts of any length: [number of errors/total number of words] × 10.

Intra-rater reliability was established by rerating 10% of the essays 3 weeks after the initial rating using the intra-class correlation coefficient (ICC). A strong degree of correlation was found, *ICC* = .99, 95% CI [.96, .99], *p* < .001. The same procedure was used to establish inter-rater reliability. An experienced colleague rated 10% of the essays. The error ratios were correlated with those of this study's rater. This also found a high level of agreement, *ICC* =.99, 95% CI [.91, .99], *p* < .001.

Data Analysis Procedures

First, the error ratios were calculated for the first draft of each of the four writing tasks and then compared. The two factors were time (i.e. essay) and condition (i.e. group). Time was a repeated-measures within-group independent variable, and condition was a between-group independent variable. The dependent variable was error ratios.

The assumptions for a parametric ANOVA were not met. One issue was the sample size (*N* = 30). Additionally, although Levene's tests showed that homoscedasticity was satisfied for the between-group factor, the data violated the assumption of sphericity for the repeated measures factor. Mauchley's test indicated that for the indirect-feedback group (*W* = 0.11, *p* = .005) and the direct-feedback group (*W* = 0.22, *p* = .024) the condition of sphericity was violated, which is telling given that the test lacks power for small sample sizes (Field et al., 2012). Furthermore, a series of Shapiro-Wilk tests of normality by group for each essay showed that nearly half of the data subsets differed significantly from a normal distribution.

Because no non-parametric equivalent of a two-way mixed ANOVA exists (Field et al., 2012), a robust mixed ANOVA with trimmed means and bootstrapping was used. Robust tests are resistant not only to non-normal distribution, but also to the loss of information inherent to traditional non-parametric testing (Larson-Hall, 2010; Larson-Hall & Herrington, 2009). There is also evidence that bootstrapping controls errors better than more traditional adjustments (Berkovits et al., 2000). This necessitated three tests using the sppba, sppbb and sppbi functions in the WRS2 package (Mair et al., 2015) within the R software environment (R Core Team, 2016).

A final consideration was effect sizes. This was considered particularly important for this study because of its sample size. Not only are effect sizes less susceptible to the issues presented by smaller samples (Clark-Carter, 2003), but they can point to an effect that inferential testing fails to detect.

Results

Descriptive Statistics

Descriptive statistics for error ratio by session, split by group, are shown in Table 4. Trimmed means were automatically generated and applied by the software (Mair et al., 2015). Two thousand bootstrap samples were taken. Initially α was set at .05, but then adjusted using a Bonferroni correction to account for the fact that three tests were used. Final α was set at .016.

Table 4
Descriptive Statistics for Total Error Ratio by Group

Group	Session	n	Minimum error ratio	Maximum error ratio	Mean error ratio	*SD*
Control	Pretest	10	0.51	2.25	1.25	0.53
	Treatment	10	0.65	2.98	1.45	0.76
	Posttest	10	0.57	2.87	1.30	0.79
	Delayed posttest	10	0.70	2.29	1.28	0.49
Indirect WCF	Pretest	11	0.30	1.86	1.07	0.49
	Treatment	11	0.43	2.52	1.02	0.58
	Posttest	11	0.33	2.98	1.11	0.72
	Delayed posttest	11	0.24	2.86	1.03	0.69
Direct WCF	Pretest	9	0.39	2.23	1.15	0.64
	Treatment	9	0.42	2.33	1.16	0.68
	Posttest	9	0.29	2.14	1.08	0.61
	Delayed posttest	9	0.70	2.36	1.39	0.56

Inferential Statistics

Table 5 shows the main effect for group only on error ratio. It indicates whether the three groups' comparative error rates differed over the entirety of the study. At $p=.46$, the test was well short of significance. The null hypothesis that mean error ratios were not affected by treatment type cannot be rejected.

Table 5
Main Effect for Group

DFn	DFd	Comparison (Group × Group)	Ψ	*p*
		Control × Indirect	0.22	
2	27	Control × Direct	0.11	.46
		Indirect × Direct	-0.12	

Table 6
Main Effect for Time (Essay)

| DFn	DFd	Comparison (Essay × Essay)	Ψ	*p*
		1 × 2	-0.08	
		1 × 3	0.07	
		1 × 4	<0.01	
3	81	2 × 3	0.09	.49
		2 × 4	-0.03	
		3 × 4	-0.07	

Table 6 shows the main effect for essay. This shows the level of significance in changes in error ratios across time, irrespective of treatment group. The intention of this test was to ensure that error ratios were not affected by variation in difficulty between the four writing tasks. The assumption was that if error rates for any one essay varied from the others, and if this persisted across groups, it would point to inconsistent task difficulty. The results were not statistically significant (*p*=.49). There was no significant variation between any combination of pretest, essay two, posttest, or delayed posttest, suggesting no problematic variation in difficulty between the four writing tasks.

Table 7 shows the combined effect for group and time, and so reflects both within-groups and between-groups measures. The results for this test were also not significant (*p*=.69). The null hypothesis cannot therefore be rejected: There was no difference in change in accuracy over time between any of the three groups, and none of the groups improved significantly over time. Each of the three groups is represented visually in Figure 1.

Table 7

Interaction Effect for Group × Time

DFn	DFd	Group	Interaction (Group/Essay × Group/Essay)	Ψ	*p*
6		Control	1/1 × 1/2	-0.15	
			1/1 × 1/3	0.12	
			1/1 × 1/4	0.27	
			1/2 × 1/3	-0.12	
			1/2 × 1/4	-0.02	
			1/3 × 1/4	0.10	
	81	Indirect WCF	2/1 × 2/2	-0.24	.69
			2/1 × 2/3	0.22	
			2/1 × 2/4	0.45	
			2/2 × 2/3	0.19	
			2/2 × 2/4	-0.08	
			2/3 × 2/4	-0.27	
		Direct WCF	3/1 × 3/2	0.11	
			3/1 × 3/3	0.27	
			3/1 × 3/4	0.16	
			3/2 × 3/3	-0.20	
			3/2 × 3/4	0.11	
			3/3 × 3/4	0.31	

Figure 1
Error Ratios over Time by Group
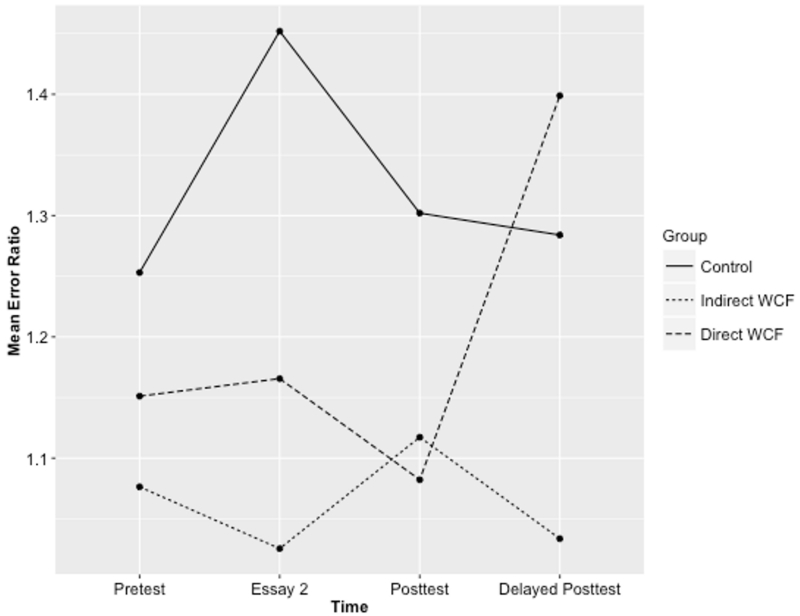
Note. Lower values denote greater accuracy.

Effect Sizes

The between-groups effect sizes are presented in Table 8. A visual representation of these contrasts is given in Figure 2. These data show effect sizes for the between-group comparisons for each essay. Based on Cohen's (1988) effect size benchmarks, there was only one pattern that might point to a notable effect; there was a consistent difference in means in favour of the indirect WCF group over the control group. This was manifested in a small effect for the pretest (d=0.35), a medium effect for essay 2 (d=0.63), a small effect for the posttest (d=0.25), and a small effect for the delayed posttest (d=0.42). This could be construed as indirect WCF having a positive effect on accuracy over no WCF at all. However, the fact that this effect was present in the pretest, and therefore preceded any treatments, suggests that it is more likely to be indicative of a disparity in proficiency between the two groups that was not detected by earlier controls.

Table 8

Effects Sizes for Between-Group Comparisons by Time

Time (Essay)	Comparison (Groups)	d	95% CI
1	Control×Indirect WCF	0.35	-0.57, 1.27
	Control×Direct WCF	0.17	-0.77, 1.12
	Indirect WCF×Direct WCF	0.14	-1.09, 0.80
2	Control×Indirect WCF	0.63	-0.31, 1.57
	Control×Direct WCF	0.40	-0.55, 1.35
	Indirect WCF×Direct WCF	0.22	-1.19, 0.75
3	Control×Indirect WCF	0.25	-0.67, 1.17
	Control×Direct WCF	0.31	-0.64, 1.26
	Indirect WCF×Direct WCF	0.04	-0.92, 1.01
4	Control×Indirect WCF	0.42	-0.50, 1.35
	Control×Direct WCF	-0.21	-1.16, 0.74
	Indirect WCF×Direct WCF	-0.57	-1.56, 0.42

Note. Negative values denote overall higher accuracy for the fixed variable (i.e. the group listed first in the contrast). CI = confidence interval.

Figure 2

Effect Sizes for Between-Group Comparisons by Time
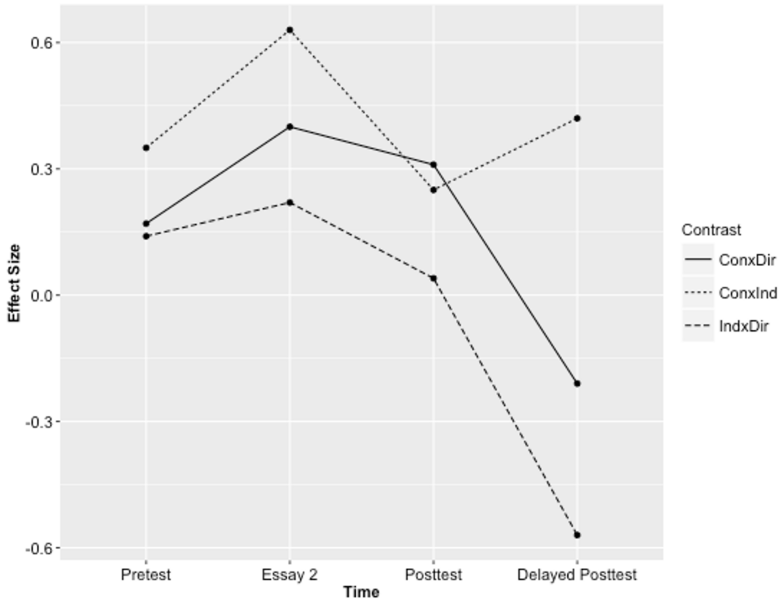
Note. ConxDir = Control×Direct WCF; ConxInd = Control×Indirect WCF; IndxDir = Indirect×Direct WCF

There were also less obvious patterns. The relationship of the direct WCF group to both of the other groups followed a similar trajectory (see Figure 2). For Essay 2 and the posttest the direct feedback group outperformed the control group (d=0.40 and d=0.31, respectively), but the reverse was true for the delayed posttest (d=-0.21). Comparing the indirect WCF and direct WCF groups shows a small effect for direct WCF over indirect for Essay 2 (d=0.22), which became trivial by the posttest (d=0.04) and then became a medium effect size in favour of the indirect WCF group in the delayed posttest (d=-0.57). These data suggest that over time both the indirect WCF group and the control group improved in accuracy, where the direct WCF did not.

However, the effect sizes for within-groups comparisons show that this cannot be the case (see Table 9). All of the effect sizes for the indirect feedback group were trivial, which suggests no change in accuracy over the course of the study. In fact, none of the groups improved, but the direct

feedback group declined between the posttest and the delayed posttest (see Figure 1). This could not be due to a gain in accuracy that was subsequently lost in the delayed posttest for two reasons. Firstly, the gains made by the direct WCF group from the pretest to Essay 2 and the posttest were trivial (d=0.02, d=0.11, respectively), so there was effectively no improvement. Secondly, the accuracy in the delayed posttest was substantially below even that of the pretest (d=-0.40). This was an actual decline in accuracy, not a decline relative to earlier gains.

Table 9
Effect Sizes for Within-Group Comparisons by Time

Group	Comparison (Time)	d	95% CI
Control	1×2	-0.31	-1.25, 0.64
	1×3	-0.07	-1.01, 0.87
	1×4	-0.06	-1.00, 0.88
	2×3	0.19	-0.75, 1.14
	2×4	0.27	-0.68, 1.21
	3×4	0.03	-0.91, 0.97
Indirect WCF	1×2	0.09	-0.80, 0.98
	1×3	-0.06	-0.95, 0.82
	1×4	0.07	-0.82, 0.96
	2×3	0.14	-1.03, 0.75
	2×4	0.02	-0.91, 0.87
	3×4	0.11	-0.78, 1.00
Direct WCF	1×2	0.02	-1.01, 0.98
	1×3	0.11	-0.89, 1.11
	1×4	-0.40	-1.38, 0.58
	2×3	0.12	-0.88, 1.12
	2×4	-0.37	-1.38, 0.61
	3×4	-0.53	-1.55, 0.49

Note. Negative values denote an overall decline in accuracy over time. CI = confidence interval.

The effect sizes for the control group were similarly influenced by a single unusual result, not by a pattern of improvement or decline. Accuracy declined in essay two for the control group. This led to small positive effect sizes from Essay 2 to the posttest (d=0.19) and the delayed posttest (d=0.27). Again, however, this is due to lower accuracy in one essay, not to a pattern of improvement. Effect sizes show a trivial decrease in accuracy for the control group if the pretest is compared to the posttest (d=-0.07) and the delayed posttest (d=-0.06). There is no pattern of change in accuracy in the control group.

Effect sizes for this study, therefore, accord with the inferential statistics. There is no pattern of change between any of the three groups. The null hypothesis cannot be rejected for any of the research questions investigated in this study.

Discussion

The results of this study differ from much of the published research. One possible reason for this is the sample. Although a sample size of 30 participants is not unusual in applied linguistics research, this was compounded by the division of the sample into three smaller groups. It is therefore possible that the results of this study are due to type II error. However, it is unlikely that this is the case for two reasons. The first is the use of robust statistical tests, which have been shown to increase power with small, nonrandom samples (Wilcox, 1998, 2010). More significant are the effect sizes reported in Tables 8 and 9. Unlike inferential statistics, effect sizes are not particularly sensitive to sample size (Clark-Carter, 2003). There is nothing in the effect sizes to suggest any change in written accuracy for any of the three groups. Although each of the confidence intervals taken in isolation is not compelling, the fact that such a large number of effect sizes cluster around trivial to small values paints a clearer picture. The question of whether this study generated type II errors therefore becomes academic. Even if inferential statistics had uncovered a statistically significant effect, it would have been rendered inconsequential by the trivial effect sizes.

A further possible reason the study did not reproduce the results of recent studies is that it applied comprehensive WCF. However, two of the better studies of WCF (Van Beuningen et al., 2008, 2012) focused on comprehensive feedback and found an effect. As long as students are given sufficient time to process and act on comprehensive feedback, there seems to be no reason to think it is inherently ineffective.

There is a final possibility that speaks more directly to the aims of this study. What separates this study from the majority of published studies may not be methods, but context. There are three salient points here: The study took place in a Japanese EFL context, it was not classroom-based, and the participants were not currently studying English.

A key factor here that distinguishes the current study from much of the published literature is its setting. The evidence for the efficacy of WCF in EFL settings is mixed. The current study, as well as Robb et al. (1986) and Truscott and Hsu (2008) all dealt with EFL populations and found no effect for WCF. Ellis et al. (2008) found modest effects for WCF, but they focussed only on article use. Farrokhi and Sattarpour (2012), on the other hand, found an effect for WCF with their "high-proficient" Iranian EFL students. This pattern of results suggests that the ways in which EFL and ESL populations typically differ may provide some explanation for the outcomes of the current study. One possibility is that if we accept that students in ESL contexts tend to benefit from a higher degree of automaticity in L2 processing, then it may be that they have more resources available to deal with the cognitive demands of uptake of WCF. Farrokhi and Sattarpour's participants may have fared better because of their higher degree of proficiency, something they share with most ESL students.

There is also the question of what effect classroom input has on WCF uptake. This is something upon which the current study is positioned to offer a unique perspective. This study was designed to control for language input. In order to do this, only participants who were not currently studying English in any formal sense were enlisted. This, combined with the EFL setting, meant that most of the participants experienced little exposure to English over the course of the study. In other words, the current study is as close as we have come to a genuinely experimental study of the effects of WCF on written accuracy. From this perspective, the fact that the results of the study were negative is revealing. It hints at the possibility that WCF alone is ineffective, and that what we see in positive studies is the result of input, not correction. In other words, the outcomes of more positive studies may be confounded by external exposure to English while the study was conducted. The use of control groups has probably mitigated this issue somewhat, but without further research we are not currently in a position to rule it out.

A final possibility is that the experimental nature of the study led participants to attend to WCF in a perfunctory manner, and so it was ineffective. Bruton (2009, 2010) has argued that a key component in the

uptake of WCF is "such factors as . . . grades [that] will make a difference and cannot be ignored" (2010, p. 496). This argument has not yet been substantiated by research, but it is plausible. We can speculate that students who attend to WCF outside the stakes, pressures, and social dynamics of the classroom may fail to do so substantively.

Conclusion

It would be a mistake to suggest abandoning WCF based only on the results of the current study. One issue that militates against this is the sample size. This said, the effect sizes showed that the non-significant findings were very probably not due to the small sample. Had the results of significance testing and effect sizes diverged, this would have pointed to a need to replicate the study with a larger sample-size. However, both showed no effect for either form of WCF.

Ecological validity was compromised by the fact that this study was experimental research that took place outside the classroom. Data derived from such controlled conditions need to be treated with caution. It should be noted however that it was exactly the degree of control over input exercised in this study that made it unique and was responsible for its more interesting possible implications.

An additional limitation of the study is its length. This poses the question of whether we should expect to see substantive improvement in written accuracy over a period of 2 months. This is not a limitation of this study alone, but of most research into the efficacy of WCF. In fact, some recent studies have included only a single treatment over a considerably shorter period.

The study can however offer tentative recommendations within a limited context. From a pedagogical perspective the findings arguably have more to say to teachers in a Japanese EFL context than to writing teachers more generally. Such teachers should be open to the possibility that WCF may be ineffective. The decision of whether to engage students in the time-consuming cycle of error correction and revision should be based on two considerations. First, how many times will I offer feedback to my students? This study suggests that once or twice will be insufficient for any effect. Second, are my students ready to grapple with the demands of this task and to learn from it? This may be beyond them if they already struggle with the cognitive burden of writing in English.

The current study also points to further research that could inform improved teaching practices in this area. The efficacy of WCF for EFL

populations is not well understood. Evidence in these contexts is weak. This suggests that different populations may respond to WCF differently, and the research should reflect this if it is to inform teaching outside a limited context. It is important that we understand not only how EFL and ESL populations differ, but how these differences may influence the uptake of WCF.

The current study was unique in the control it exercised over language input as a confounding factor. Although this imposed limitations, it also helped to identify an area about which we understand little, but which may be important in informing classroom practices. That this study found no effect for WCF suggests that external exposure to English may be a factor in some of the positive evidence for WCF. Research that attempts to establish the extent to which this is true could be invaluable to EFL instructors and their students.

Statistically, the results of this study were null. No effect on accuracy was found for either form of feedback. The negative results were unexpected, but this study has a contribution to make to both pedagogy and possible directions for future research. It was suggested that teachers who share the study's Japanese EFL setting should approach WCF with caution and consider whether their context is likely to be conducive to student uptake of such feedback.

Darby McGrath is a tenured senior lecturer at Ritsumeikan Asia Pacific University. His primary research interest is academic writing, specifically the use of grammatical metaphor.

References

Adams, R. (2003). L2 output, reformulation and noticing: Implications for IL development. *Language Teaching Research, 7*(3), 347-376. https://doi.org/10.1191/1362168803lr127oa

Berkovits, I., Hancock, G. R., & Nevitt, J. (2000). Bootstrap resampling approaches for repeated measures designs: Relative robustness to sphericity and normality violations. *Educational and Psychological Measurement, 60*(6), 877-892. https://doi.org/10.1177/ 00131640021970961

Bitchener, J. (2008). Evidence in support of written corrective feedback. *Journal of Second Language Writing, 17*(2), 102-118. https://doi.org/10.1016/j.jslw.2007.11.004

Bitchener, J. (2012). A reflection on 'the language learning potential' of written CF. *Journal of Second Language Writing, 21*(4), 348-363. https://doi.org/10.1016/j.jslw.2012.09.006

Bitchener, J., & Ferris, D. (2012). *Written corrective feedback in second language acquisition and writing*. Routledge. https://doi.org/10.4324/9780203832400

Bitchener, J., & Knoch, U. (2008). The value of written corrective feedback for migrant and international students. *Language Teaching Research, 12*(3), 409-431. https://doi.org/10.1177/1362168808089924

Bitchener, J., & Knoch, U. (2009). The value of a focused approach to written corrective feedback. *ELT Journal, 63*(3), 204-211. https://doi.org/10.1093/elt/ccn043

Bitchener, J., & Knoch, U. (2010a). Raising the linguistic accuracy level of advanced L2 writers with written corrective feedback. *Journal of Second Language Writing, 19*(4), 207-217. https://doi.org/10.1016/j.jslw.2010.10.002

Bitchener, J., & Knoch, U. (2010b). The contribution of written corrective feedback to language development: A ten month investigation. *Applied Linguistics, 31*(2), 193-214. https://doi.org/10.1093/applin/amp016

Bitchener, J., Young, S., & Cameron, D. (2005). The effect of different types of corrective feedback on ESL student writing. *Journal of Second Language Writing, 14*(3), 191-205. https://doi.org/10.1016/j.jslw.2005.08.001

Bruton, A. (2009). Designing research into the effects of grammar correction in L2 writing: Not so straightforward. *Journal of Second Language Writing, 18*(2), 136-140. https://doi.org/10.1016/j.jslw.2009.02.005

Bruton, A. (2010). Another reply to Truscott on error correction: Improved situated designs over statistics. *System, 38*(3), 491-498. https://doi.org/10.1016/j.system.2010.07.001

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*(3), 267-296. https://doi.org/10.1016/S1060-3743(03)00038-9

Clark-Carter, D. (2003). Effect size: The missing piece in the jigsaw. *The Psychologist*, *16*(12), 636-638.

Cohen, A. D., & Robbins, M. (1976). Toward assessing interlanguage performance: The relationship between selected errors, learners' characteristics, and learners' explanations. *Language Learning, 26*(1), 45-66. https://doi.org/10.1111/j.1467-1770.1976.tb00259.x

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Erlbaum Associates.

Ellis, R., Sheen, Y., Murakami, M., & Takashima, H. (2008). The effects of focused and unfocused written corrective feedback in an English as a foreign language context. *System, 36*(3), 353-371. https://doi.org/10.1016/j.system.2008.02.001

Farrokhi, F. & Sattarpour, S. (2012). The effects of direct written corrective feedback on improvement of grammatical accuracy of high-proficient L2 learners. *World Journal of Education, 2*(2), 49-57. https://doi.org/10.5430/wje.v2n2p49

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* (pp. 81-104). Cambridge University Press.

Ferris, D., Liu, H., Sinha, A., & Senna, M. (2013). Written corrective feedback for individual L2 writers. *Journal of Second Language Writing, 22*(3), 307-329. https://doi.org/10.1016/j.jslw.2012.09.009

Field, A. P., Miles, J., & Field, Z. (2012). *Discovering statistics using R*. Sage.

Frear, D., & Chiu, Y. (2015). The effect of focused and unfocused indirect written corrective feedback on EFL learners' accuracy in new pieces of writing. *System, 53*, 24-34. https://doi.org/10.1016/j.system.2015.06.006

Karim, K., & Nassaji, H. (2018). The revision and transfer effects of direct and indirect comprehensive corrective feedback on ESL students' writing. *Language Teaching Research, 24*(4), 519-539. https://doi.org/10.1177/1362168818802469

Kepner, C. G. (1991). An experiment in the relationship of types of written feedback to the development of second-language writing skills. *The Modern Language Journal, 75*(3), 305-313. https://doi.org/10.1111/j.1540-4781.1991.tb05359.x

Lalande, J. F., II. (1982). Reducing composition errors: An experiment. *The Modern Language Journal, 66*(2), 140-149. https://doi.org/10.2307/326382

Larson-Hall, J. (2010). *A guide to doing statistics in second language research using SPSS*. Routledge.

Larson-Hall, J., & Herrington, R. (2009). Improving data analysis in second language acquisition by using modern developments in applied statistics. *Applied Linguistics, 31*(3), 368-390. https://doi.org/10.1093/applin/amp038

Mair, P., Schoenbrodt, F., & Wilcox, R. R. (2015). WRS2: Wilcox robust estimation and testing (Version 0.4-0) [Computer software].

Mair, P., Wilcox, R. R., & Schoenbrodt, F. (2015). Package "WRS2." https://cran.r-project.org/web/packages/WRS2/WRS2.pdf

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*(1), 43-68. https://doi.org/10.1016/S1060-3743(98)90005-4

Qi, D. S., & Lapkin, S. (2001). Exploring the role of noticing in a three-stage second language writing task. *Journal of Second Language Writing, 10*(4), 277-303. https://doi.org/10.1016/S1060-3743(01)00046-7

R Core Team. (2016). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing (Version 3.3.2) [Computer software]. https://www.R-project.org

Robb, T., Ross, S., & Shortreed, I. (1986). Salience of feedback on error and its effect on EFL writing quality. *TESOL Quarterly, 20*(1), 83-95. https://doi.org/10.2307/3586390

Sachs, R., & Polio, C. (2007). Learners' uses of two types of written feedback on a L2 writing revision task. *Studies in Second Language Acquisition, 29*(1), 67-100. https://doi.org/10.1017/S0272263107070039

Semke, H. D. (1984). Effects of the red pen. *Foreign Language Annals, 17*(3), 195-202. https://doi.org/10.1111/j.1944-9720.1984.tb01727.x

Sheen, Y. (2007). The effect of focused written corrective feedback and language aptitude on ESL learners' acquisition of articles. *TESOL Quarterly, 41*(2), 255-283. https://doi.org/10.1002/j.1545-7249.2007.tb00059.x

Sheen, Y., Wright, D., & Moldawa, A. (2009). Differential effects of focused and unfocused written correction on the accurate use of grammatical forms by adult ESL learners. *System, 37*(4), 556-569. https://doi.org/10.1016/j.system.2009.09.002

Sheppard, K. (1992). Two feedback types: Do they make a difference? *RELC Journal, 23*(1), 103-110. https://doi.org/10.1177/003368829202300107

Shintani, N., & Ellis, R. (2013). The comparative effect of direct written corrective feedback and metalinguistic explanation on learners' explicit and implicit knowledge of the English indefinite article. *Journal of Second Language Writing, 22*(3), 286-306. https://doi.org/10.1016/j.jslw.2013.03.011

Storch, N., & Wigglesworth, G. (2010). Learners' processing, uptake, and retention of corrective feedback on writing. *Studies in Second Language Acquisition, 32*(2), 303-334. https://doi.org/10.1017/S0272263109990532

Truscott, J. (1996). The case against grammar correction in L2 writing classes. *Language Learning, 46*(2), 327-369. https://doi.org/10.1111/j.1467-1770.1996.tb01238.x

Truscott, J. (1999). The case for "The case against grammar correction in L2 writing classes": A response to Ferris. *Journal of Second Language Writing, 8*(2), 111-122. https://doi.org/10.1016/S1060-3743(99)80124-6

Truscott, J. (2004). Evidence and conjecture on the effects of correction: A response to Chandler. *Journal of Second Language Writing, 13*(4), 337-343. https://doi.org/10.1016/j.jslw.2004.05.002

Truscott, J. (2007). The effect of error correction on learners' ability to write accurately. *Journal of Second Language Writing, 16*(4), 255-272. https://doi.org/10.1016/j.jslw.2007. 06.003

Truscott, J. (2009). Arguments and appearances: A response to Chandler. *Journal of Second Language Writing, 18*(1), 59-60. https://doi.org/10.1016/j.jslw.2008.09.001

Truscott, J. (2010). Some thoughts on Anthony Bruton's critique of the correction debate. *System, 38*(2), 329-335. https://doi.org/10.1016/j.system.2010.03.014

Truscott, J., & Hsu, A. Y. (2008). Error correction, revision, and learning. *Journal of Second Language Writing, 17*(4), 292-305. https://doi.org/10.1016/j.jslw.2008.05.003

Van Beuningen, C., De Jong, N. H., & Kuiken, F. (2008). The effect of direct and indirect corrective feedback on L2 learners' written accuracy. *ITL - International Journal of Applied Linguistics, 156*, 279-296. https://doi.org/10.2143/ITL.156.0.2034439

Van Beuningen, C., De Jong, N. H., & Kuiken, F. (2012). Evidence on the effectiveness of comprehensive error correction in second language writing. *Language Learning, 62*(1), 1-41. https://doi.org/10.1111/j.1467-9922.2011.00674.x

Wilcox, R. R. (1998). How many discoveries have been lost by ignoring modern statistical methods? *American Psychologist, 53*(3), 300-314. https://doi.org/10.1037/0003-066X.53.3.300

Wilcox, R. R. (2010). *Fundamentals of modern statistical methods: Substantially improving power and accuracy* (2nd ed.) Springer.

Appendix A
Correction Codes for Indirect Feedback Group

Symbols!	Examples!	Corrections!
o = omit	**o** I work <u>for</u> to make money.	*I work to make money.*
sp = spelling!	**sp** for two <u>mounths</u>	***months***!
ww = wrong word!	**ww** I spent a lot of money <u>for</u> my car **ww** Please <u>teach</u> me your email address.	***on*** ***tell!***
sv = subject verb agreement!	**sv** She <u>watch</u> television every night.	*She **watches …**!*
vt = verb tense!	**vt** I <u>work</u> after school yesterday.	*I **worked …**!*
wf = word form!	**wf** I am <u>interesting</u> in English. **wf** My hometown is a very <u>safety</u> place.	*I am **interested** in …* *… a very **safe** place.**!***
∧ = missing Word or words !	I love listening⸦music. I have liked it since⸦twelve years old.	*… listening **to** music* *… since **I was** twelve years old!*
# = change singular or plural !	**#** I like reading <u>book</u>. **#** I'd like to be a <u>teachers</u>.	*… reading **books**.* *...a **teacher**.!*
△ = article mistake	I work in⸦store in Osaka. I visited ⸦an Eiffel Tower<u>.</u>	*… in **a** store …* *… **the** Eiffel Tower.*
c = Capital Letter Mistake	**c** I didn't go to school on <u>monday</u>**.** **c** My favourite film is <u>HARRY POTTER</u>.	***Monday*** ***Harry Potter***
p = punctuation	**p** <u>Also baseball</u> is an interesting sport.	***Also, baseball…***
wo= word order	**wo** I caught a <u>fish very big</u>.	***… a very big fish.***
frg=fragment *(not a complete sentence)*	**frg** I love playing sport. <u>For example, football and tennis</u>. Having brothers and sisters is **frg** helpful. <u>Because they can give me advice.</u>	*I love playing **sport, for example** football and tennis.* *Having brothers and sisters is helpful **because** they can give me advice.*

Appendix B

Partial Example of Indirect WCF

"Essay 3, Draft 1"

There are many features of a good friend. The first feature is being funny and

interesting. It is easy to understand why this is a feature of a good friend P because you

would not want to be with a person that has no humor and has no interest in you. Also,

being with a person that is funny would also make ꙍ yourself happy and comfortable. The

second feature is being respective. There is a saying in Japan, "Be respective, even to

the closest friend". This shows that even close friends need to respect each other in

order to maintain a good relationship.

Appendix C

Partial Example of Direct WCF

"Essay 3, Draft 1"

There are many features of a good friend. The first feature is being funny and

interesting. It is easy to understand why this is a feature of a good friend, because you

would not want to be with a person that has no humor and has no interest in you. Also,

being with a person that is funny would also make ~~yourself~~ you happy and comfortable. The

second feature is being ~~respective~~ respectful. There is a saying in Japan, "Be ~~respective~~ respectful, even to

the closest friend". This shows that even close friends need to respect each other in

order to maintain a good relationship.