Issues in Language Teaching (ILT), Vol. 7, No. 1, 87-110, June 2018

# Motivating the Unmotivated: Making Teacher Corrective Feedback Work

**Masoud Azizi***
*Assistant Professor, Amirkabir University of Technology*

**Majid Nemati**
*Associate Professor, University of Tehran*

**Abstract**
It is often wrongly assumed that the provision of teacher corrective feedback naturally entails learners' attendance to and application of it, but learners have repeatedly been reported not to pay attention to teacher feedback due to lack of motivation and the distracting effect of the grades they receive. The present study was an attempt to tackle this problem. To do so, the technique named Draft-Specific Scoring (Nemati & Azizi, 2013) was implemented. In DSS, learners receive both teacher feedback and grades on their first drafts; however, they are given up to two opportunities to apply teacher feedback and revise their drafts accordingly. The scores they receive may improve as a result of the quality of revisions they make. Students' final score will be the mean score of all the grades they receive on the final drafts of their assignments. For this purpose of the present study, 57 Iranian intermediate students attending the 'Advanced Writing' course at University of Teheran, with an age range of 21 to 27 were studied in two groups. The gain score analysis and the SPANOVA used showed the superiority of DSS over more traditional methods in improving learners' overall writing proficiency as well as fluency and accuracy of their written texts. Moreover, no adverse effect was observed for the treatment group regarding the grammatical complexity of their texts. This indicates that in order to make teacher feedback work, there are a number of intervening variables one needs to consider, the most important of which being learners' motivation to attend to teacher feedback.

***Keywords***: Corrective feedback, grading; draft specific scoring, fluency, accuracy, grammatical complexity

**Corresponding author:** Masoud Azizi (mazizi@aut.ac.ir, nematim@ut.ac.ir)

## INTRODUCTION

In case learners do not attend to or engage with teacher corrective feedback, it would be very difficult to be able to comment on the effectiveness or ineffectiveness of corrective feedback (Nemati & Azizi, 2013, Azizi & Nemati, 2018). If a teacher comments on a learner's piece of writing but keeps it in her own drawer, she cannot call her comments 'teacher feedback.' Teachers' comments may be regarded as feedback only when learners have access to them and engage in an interaction with the given feedback by processing and applying it to their writings. If one of these elements is missing, it does not seem plausible to regard teachers' comments cannot be regarded as feedback.

However, something that has been neglected in the debate on the effectiveness of corrective feedback is the underlying assumptions of Truscott's (1996) thesis and the whole debate on the ineffectiveness of corrective feedback. What is implied in that debate is that corrective feedback includes teacher's provision of comments PLUS learners' attendance to and probably the application of those comments.

Over the past 3 decades, scholars have been arguing about the effectiveness of corrective feedback as if they had ensured the presence of all these elements in all the studies examining teacher feedback. In fact, in all such debates, only the provision of feedback was present but no information was available on the presence or quality of the other elements, i.e., learners' attendance to and application of teacher feedback. As a result, arguing about the effectiveness or ineffectiveness of corrective feedback does not seem plausible without first ensuring the presence of all the necessary elements.

In addition, the myriad of conflicting results in the literature regarding the effectiveness of corrective feedback (Bitchener & Ferris, 2012; Lee, 2014; Mawlawi-Diab, 2015; Zheng & Yu, 2018) is the evidence for the fact that learners' attendance has not been ensured or at least checked in the majority of studies on corrective feedback so far.

## LITERATURE REVIEW

After Truscott (1996) published his paper questioning the value of grammar correction in writing classes, a myriad of reactions emerged in all writing

journals regarding the validity of Truscott's thesis. He claims that grammar correction is not only ineffective but also harmful because it can hinder the learning process. He believes that teachers do it because they simply think it should be effective without having any basis for their belief. Truscott (1996) believes that learners are often not motivated enough to attend to teacher feedback, and even if they do, they may not be motivated enough to apply it to their writings. On the other hand, he believes that learners who do not receive correction often enjoy a more positive attitude toward writing. He believes that the time spent on correction by teachers and students can be better invested in other aspects of writing. He argues that it is not only acceptable for teachers not to correct learners' errors, but it is also in fact preferable for them not to do so. He claims that grammar correction should have no place in the realm of writing instruction and as a result should be abandoned.

Truscott (2007), doing a meta-analysis on the studies carried out on corrective feedback, claims that corrected students may shorten or simplify their writings in order to avoid situations in which they are more likely to make mistakes and be corrected as a result. In other words, corrected students will try to hide their weaknesses. Therefore, where their scores in overall accuracy improve, it might simply be due to the fact that they have found a way to avoid structures they are not sure of. Accordingly, he concludes that avoidance may have biased the results in favor of error correction both in case of absolute gains and in relation to uncorrected students.

Ferris (1999), having called Truscott's anti-correction thesis "premature and overly strong," (p. 2) argues that Truscott's definition of error correction is vague and problematic. She believes that "selective, prioritized, and clear" (p. 4) error correction can and does help at least some students.To Ferris, for effective grammar feedback and instruction, one should consider issues such as learners' first language background, their English language proficiency level, and their prior experience with grammar instruction and editing strategies. Moreover, teachers need to raise learners' motivation by making them aware of the importance of accuracy in their written texts, and they need to develop independent self-editing skills.

Ferris (1999), unlike Truscott, argues that we should continue error correction in L2 because surveys indicate that learners highly value and

demand teacher feedback, and the absence of any form of grammar feedback may frustrate the writing class especially when learners observe that according to the scoring rubrics and proficiency tests, their language errors can prevent them from achieving success in their educational and professional life.

Ferris (2004) believes that the large variation in the research designs of the studies carried out before on corrective feedback makes it very difficult to draw any conclusion regarding the effectiveness or ineffectiveness of corrective feedback. She states that they varied on every research parameter including subject characteristics, sample size, treatment duration, the type of writing being considered, the type of feedback being given, the person providing the feedback, how errors were defined, and how improvement and accuracy were assessed. Guenette (2007) also agrees with Ferris on this point. Ferris (1999, p. 9) believes that only when answers to the following and similar questions were sought systematically, one can definitely support or refute Truscott's thesis.

- Which individual student variables affect learners' willingness and ability to benefit from error correction, and can student problems be mitigated by thoughtful pedagogical practice?
- Which methods, techniques, or approaches to error correction lead to short- or long-term student improvement (assuming that student, teacher, and contextual variables are adequately controlled for)?

Bruton (2009), objecting to Truscott's anti-correction position, asserts that common sense and intuition entails that correction cannot be harmful to developing accuracy; moreover, it does not seem plausible to assume that lack of correction or simply more writing practice can be conducive to improvement. To him, it is both logical and intuitive that more evidence, either positive or negative, results in improvement in learners' level of correctness, "unless the evidence is incomprehensible, erroneous, confusing, or just conflicting, thus causing backsliding" (p. 604).

Bruton (2010) also emphasizes on the relationship between motivation and effort to improve. He believes that factors such as instruction, tasks, and grades can affect learners' success and should not be overlooked. He explains that often the participants are not given any purpose or objective for what they are supposed to do, and sometimes no feedback on content or encouragement is given to students in L2 writing research. In other cases, no

grades are provided or if grades are given, no reference is made to content, which encourages avoidance. All these can demotivate learners. He believes that students need to have a reason for trying to improve their accuracy level. To him, the climate of the response as well as grades is so important in this regard. However, Truscott (2010) criticizes Bruton (2010) for not presenting any study in which motivation was present and correction was found helpful.

Bruton's (2010) belief in the role of grades seems intriguing, but the literature on grading learners' writing indicates that this practice has its own flaws and problems. Grades can divert learners' attention away from teacher feedback. Students have been frequently observed ignoring teacher feedback when they see a grade on their paper (Lee, 2009).

Although instructors are aware of the harm grading can cause to learners (Lee, 2009), they continue doing so because to some extent they are required to. When teaching a writing course, most of the time we have to assign a grade to each learner at the end of the semester. This summative evaluation is what most educational institutes require their teachers to do. Scoring learners' writing samples during a semester can help have a better assessment of their writing ability at the end of the course.

Teachers also strongly believe in the role of the grades. Li and Barnard (2011), studying tutors responding to and commenting on students' writings, sought the extent to which their participants attached importance to the awarding of grades when giving feedback. In the interviews, all participants considered awarding a grade as an integral part of the feedback. One interviewee remarked that he gave feedback because it would help students get a better score. Another one said that written feedback could explain how and why a student got a certain grade. Li and Barnard (2011, p. 146) argue that according to their findings, tutors' main reason in providing learners with feedback was "less that of seeking to improve the students' writing skills and more that of justifying – to themselves, to their students, and to their academic superiors – the award of a specific grade for the assignments to hand."

Moreover, learner engagement with teacher corrective feedback has been found to be dynamic and vary across individuals (Zheng & Yu, 2018), which is affected and mediated by both learner factors and contextual factors simultaneously (Han, 2019). Learners' beliefs can have a tremendous effect on their engagement with teacher CF, for instance (Han, 2017). Students have

been frequently reported demanding their teachers to assess their writing by assigning it a grade (Lee, 2008) mostly because its interpretation is much easier for them in comparison with the sometimes vague or excessive amount of comments written in the paper margins. "Teachers should consider students' beliefs when providing WCF, and foster the development of learner beliefs conducive to deep engagement with WFC" (Han, 2017, p. 133).

In addition, one should not overlook learners' feelings when being engaged with teacher corrective feedback as they can affect the way they interact with teacher feedback. The literature abounds with studies on learners' different emotional reactions to teacher comments (Han & Hyland, 2019; Zhang & Hyland, 2018). Some students have been reported to feel proud (Ferris, Liu, Sinha, & Senna, 2013) and self-confident (Storch & Wigglesworth, 2010), others were frustrated (Zheng & Yu, 2018), indifferent, relieved, or even excited (Han & Hyland, 2015). This indicates that whatever method of feedback provision we adopt, we need to be considerate of the feelings it may trigger in our students.

## PURPOSE OF THE STUDY

Learners often expect teachers to accompany their corrective feedback with a grade summarizing the teacher evaluation of their works. In addition, teachers feel they need to score learners' writing samples due to their institutional regulations and obligations and because of the summative nature of instructional programs. On the other hand, grading learners' writing samples is problematic. Learners often ignore teacher feedback as soon as they see the grade on their paper. However, teachers continue to grade learners' writings because they feel they need to even though they believe that it may have harmful effects. However, Learning does not take place if learners do not notice teacher feedback or do not apply it in their later writing samples due to the grades they receive. At the same time, it seems that not assessing student writing is not an option at least in most contexts. As such, what is needed is a middle ground that satisfies both students' demands for their writing samples being scored and teachers' requirements while not jeopardizing their attendance to teacher feedback.

The solution we thought of was a simple technique we named Draft-Specific Scoring or simply DSS (Azizi, 2013; Azizi & Nemati, 2018;

Nemati & Azizi, 2013). In DSS, learners are provided with both corrective feedback and a grade which represent teacher's general evaluation of their writing. On the other hand, the grades learners receive are subject to change and improvement based on the quality of the revisions they make according to the teacher feedback. Learners can improve their grades by applying teacher feedback to their writings. This improvement may also be initiated by the learner herself as a result of her reflection on the way she could improve her writing in terms of both structure and content. Often, students have two chances to go through the procedure of redrafting and revising and improve their scores accordingly. Learners' final score for the whole course would be the mean score of all the grades they have received on the final version of each assignment.

As a result, the present study was an attempt to examine the effect of this technique, as a tool to motivate learners to attend to teacher corrective feedback and neutralize the negative effect of grading learners' writing samples, on learners' overall writing proficiency, change in fluency, grammatical complexity, and accuracy. The present study can be regarded as one of few studies in which it was tried to have both motivation and teacher feedback present and then assess the effectiveness of teacher feedback, a study Truscott (2010) accuses Bruton (2010) of not being able to present an example of.

# METHOD

## Participants

Two intact groups were present in this study with a total number of 57 participants (26 in the treatment group and 31 in the control group). The participants' age in the treatment group (with 10 males and 16 females) ranged from 22 to 25. The control group also consisted of 12 male and 19 female participants with an age range of 21 to 27. They were all high intermediate (based on the results of an Oxford Quick Placement Test) undergraduate students of English Language and Literature completing their BA degree at University of Tehran. They were taking part in the 'Advanced Writing' course as part of their undergraduate curriculum. All participants were Iranian but for a Chinese female participant in the treatment group.

## Procedure

Throughout the semester, the TOEFL iBT independent writing task was used as the model of the instruction. In the first few sessions, the preliminaries of writing were discussed and instructed using model essays. The fourth session was devoted to collecting learners' writing samples as pretest. Participants had 80 minutes to plan for and write about a given topic selected from among the prompts released by ETS for the task. The samples written by the two groups were compared to ensure the comparability of the two groups using TOEFL iBT scoring rubric. No significant difference was observed, $t(55) = .11$, $p = 0.91$.

As part of the instruction during the semester, some of learners' writing samples were selected and later discussed with the whole class to comment on the weaknesses and strengths. In so doing, learners' opinion was also asked. In fact, first students expressed their opinion on how the writing was and how the author could have improved it and then the teacher commented on the topic.

Each session, the students were assigned a topic to write about at home and submit their writings the following session. The papers were collected, commented on by the teacher, and returned to the learners the following session. Participants were provided with indirect feedback on their writing samples, i.e. their grammatical mistakes/errors were underlined but not corrected. All types of errors were treated. In other words, a comprehensive approach was adopted in error treatment.

For each assignment, students also received a grade representing the teacher's general evaluation of participants' writings. These scores were based on the quality of writings and were for meeting learners' demand for receiving grades for their writings. They were not used in data analysis.

The comments participants received were mainly limited to grammatical structures. For writing style-related issues including cohesion, coherence, topic relevance and topic development, some of the more problematic samples were chosen and later discussed with the whole class to shed light on how such problems should be tackled with. The majority of class time was devoted to in-class paragraph writing and group discussion on how to improve the written paragraphs in terms of style and issues other than the grammatical structure. Error feedback was limited to the comments students received on the assignments they wrote at home.

Data collection was done during two semesters with each group undergoing instruction in a different semester as the number of Advanced Writing courses offered each semester was limited. In addition, this way the difference in instruction for the two groups could not affect each other. Moreover, the two groups were kept unaware of the fact that they were being studied so that it could not affect the way they behaved and contaminate the results as a result.

Participants in both groups were strongly recommended to revise their drafts based on the comments the teacher had provided them with. Before the instruction began, both groups were informed of the method of evaluation for the course, i.e. the fact that their final score was supposed to be the mean score of all the grades they received on the final draft of their assignments during the course. In total, each group wrote 10 assignments throughout the program including the pretest, midtest, and the post test, though they did not have the chance to revise their drafts for the last two tests. As a result, they received teacher feedback and comments on only 8 assignments during the whole semester. The post test was in fact their final exam of the course they were attending. Four weeks before their final exam, the mid test was administered. There was no instruction in the time lap between the midtest and the posttest. Only one session was held after the midtest in which the students' revisions of the previous sessions were collected, and students' questions were answered regarding the comments they felt were not clear. In a sense, the posttest could be regarded as a delayed posttest in this case. The writing prompts given to the participants in pretest, midtest, and posttest were the same for both groups.

All measures were taken to keep every variable the same for both the treatment and the control group. On the other hand, one very important variable was different for the two groups. The grades participants received in the control group were fixed. In other words, they did not change as the result of the improvements or revisions learners were supposed to make based on the teacher feedback. However, in the case of the treatment group, the scores on each draft were draft-specific, that is they could change based on the revisions made by the learners, that is, in case a learner revised her first draft based on the teacher CF or her own contemplation on the topic, her score could improve on the next draft of the same assignment and it was

this new score which was taken into account to come up with the mean score at the end of the semester.

For instance, a learner who had received 15 on her first draft of an assignment could go through the comments and feedback she was provided with on her writing and think how she could revise and improve her sample accordingly. This way she could improve her grade on that assignment as well. She could receive 17, 18, or any other score based on the quality of her revised sample. There could also be no change in the score in case the quality of the revisions made was not satisfactory. The second draft was commented on by the teacher again and returned. The learners had one more chance to undergo the same procedure and improve her score. While in the control group, learners' final score was the mean of all the scores they had received on their first drafts, the mean score for the treatment group was based on all the scores they had received on their last revisions. A score profile, like the ones below was kept for each learner in each group. The participants were also advised to keep a similar one for themselves.

| | Student Name: …… | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Assignment** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** |
| **1st Draft** | 12 | 14.5 | 17 | 16 | 15 | 18 | 18 | 16 | 18 |
| **2nd Draft** | 17.5 | 16 | 19.5 | 15 | 18 | 20 | 18.5 | 18 | |
| **3rd Draft** | 18.5 | 18 | | 18.5 | | | | 19 | |
| **Final score**: The mean score of all assignments based on the last revisions (18.67) | | | | | | | | | |

**Figure 1:** The score profile for participants in the treatment group.

| | Student Name: …… | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Assignment** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** |
| **1st Draft** | 14 | 15 | 13 | 16 | 15 | 17 | 15 | 14 | 17 |
| **Final score**: The mean score of all assignments (15.11) | | | | | | | | | |

**Figure 2:** The score profile for participants in the control group.

In order to control for the handwriting effect on raters (Briggs, 1980; Hughes, Keeling, & Yuck, 1983; Klein & Taub, 2005; Russell, 2002), all essays written by both groups in the pretest, midtest, and post test were first typed. All the mistakes, no matter what type of mistake, were typed exactly as they were written by participants. All typed essays were coded by

numbers so that it was impossible for raters to identify which essay belonged to which group or which test. A detailed record of such information for each essay was kept by the researcher, however. All essays were given to two experienced raters to be rated based on TOEFL iBT writing scoring rubric for task 2. The essays were shuffled and given to raters for rating at once so that the time factor could be controlled for.

In the coordination session with raters, however, they were asked and instructed to take one step further when rating writing samples using TOEFL iBT rubric. This rubric ranges from 0 to 5. Since the participants were all at least intermediate students, the scores they could receive could be limited to the upper band scores. This could make it very difficult for the changes made as the result of the instruction to reveal themselves or be detected. As such, the raters were asked to first decide which band score each sample belonged to. Then they were asked to divide that band score into three levels (low, mid, and high) and decide to which level that sample belonged. In other words, each band score was divided into three sub-bands. For example, band score 3 was further divided into $3^-$ (read as three minus), 3, and $3^+$ (read as three plus). This way a more precise measure of learners' writing proficiency level could be obtained. In addition, the change from pretest to posttest had more room to show itself. However, for the calculation of inter-rater reliability, the band scores were only taken into account. In data entry, $1^-$ received the score zero while $5^+$ received the highest score, which was 14.

In order to check for the changes in learners' written fluency, the number of words written, as approved by Truscott (2004) himself, was used as the measure of fluency. For checking the change in the grammatical complexity of texts written by learners over time, two measures were used: The number of dependent clauses as in Robb, Ross, and Shortreed (1986) and the ratio of the number of dependent clauses to the number of clauses (Wolfe-Quintero, Inagaki, & Kim, 1998). In the case of the measure of accuracy, the ratio of error-free T-units to the number of T-units was used as introduced as the best measure of accuracy by Wolfe-Quintero et al. (1998).

In this study, a dependent clause could be any type of adverb, adjective, or noun clauses including their reduced forms. An independent clause was a clause that could stand-alone and did not need any other clause to complete its meaning. A T-unit was an independent clause with all the

dependent clauses attached to it. In other words, every independent clause was also a T-unit (Wolfe-Quintero et al., 1998). An error-free T-unit was a T-unit that did not include any kind of error but for spelling and punctuation. All these elements were counted only for the texts written in the pretest and posttest but not the midtest.

In the present study, for measures of fluency, grammatical complexity, and accuracy, only one rater rated all writing samples. As Chandler (2003) truly comments, intra-rater reliability is more important than inter-rater reliability in studies on such measures. The lowest index of intra-rater reliability was found to be .94, which is quite acceptable.

## Data Analysis

In order to examine the effect of this technique, i.e. Draft Specific Scoring, on learners' overall writing proficiency, and the change in the fluency, grammatical complexity, and accuracy of the texts written by learners, either the gain score analysis (where the assumption of the parametic tests were not met) or a mixed between-within-subjects analysis of variance was used for data analysis.

# RESULTS

The coded writing samples were given to 2 raters to be rated using TOEFL iBT independent writing scoring rubric. Using Pearson-Product Moment Correlation Coefficient, a .94 index of inter-rater reliability was obtained, and the intra-rater reliability indices for rater one and two were .90 and .92, respectively. For the purpose of data analysis the scores given by the rater with higher intra-rater reliability were used. In case there was more than one band score difference between the two raters' scores for a single writing sample, a third rater was asked to rate the sample, and the mean of the two nearest scores was used as the final score for that particular sample.

Table 1 presents the descriptive statistics for the two groups' writing scores based on the scoring rubric for the independent writing task in TOEFL iBT. While both groups started the course almost at the same level, the improvement pattern was not the same. By the time of midtest, the treatment group could improve by 2 points (from 7.35 to 9.27) while the control group could only show an improvement of less than one point (from 7.29 to 8.06). By the time the instruction was over, the treatment group had

reached 3.5 points higher than where it had started (from 7.35 to 10.88) while the control group could gain only 1.5 points (from 7.29 to 8.77).

**Table 1**: Descriptive statistics for learners' writing scores based on TOEFL iBT scoring rubric

| | Group | N | Minimum | Maximum | Mean | Std. Deviation |
|---|---|---|---|---|---|---|
| Pretest | Treatment | 26 | 4 | 12 | 7.35 | 1.83 |
| | Control | 31 | 3 | 11 | 7.29 | 2.07 |
| Midtest | Treatment | 26 | 6 | 13 | 9.27 | 2.20 |
| | Control | 31 | 5 | 12 | 8.06 | 1.95 |
| Posttest | Treatment | 26 | 7 | 14 | 10.88 | 1.97 |
| | Control | 31 | 6 | 13 | 8.77 | 2.11 |

In order to check the existance of any significant difference between the two groups in their improvement over time, a mixed between-within subjects analysis of variance was performed. There was a significant interaction between Time and Group, Wilks' Lambda = .61, $F$ (2, 54) = 16.97, $p < .0005$, partial eta squared = .39. There was also a substantial main effect for Time, Wilks' Lambda = .21, $F$ (2, 54) = 100.82, $p < .0005$, partial eta squared = .79. In addition, the main effect for Group, comparing the effect of the intervention on the two groups, was found statistically significant, $F$ (1, 55) = 4.98, $p = .03$, partial eta squared = .39, suggesting a benefit for the treatment group. According to Cohen (1988), .01 eta squared shows small effect while .06 shows a moderate effect, and .13 represents a large effect size. The results indicate that while both groups significantly improved from pretest to posttest, the treatment group undergoing DSS could significantly outperform the control group in its improvement in writing ability. The pairwise comparisons for each group among the three time periods showed that the improvements for both groups were statistically significant in all cases, i.e., the improvement from pretest to midtest, and midtest to posttest was statistically significant for both groups.

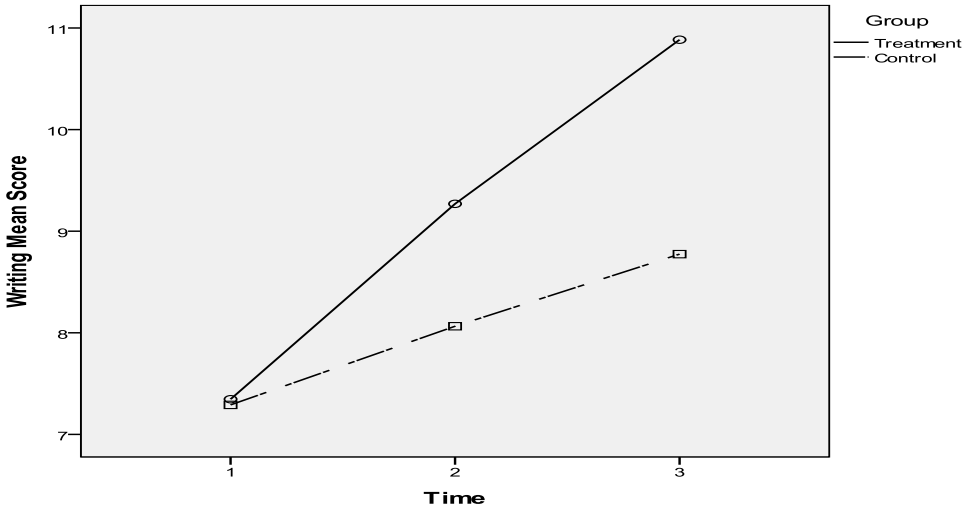


**Figure 3:** Line graph for the two groups' writing mean score across time.

Regarding changes in learners' writing fluency, both groups could improve over time in the number of words they wrote, but this improvement was more in the case of the group undergoing DSS. Table 2 presents the descriptive statistics for the two groups' mean number of words written in pretest and posttest as well as their gains.

**Table 2**: The number of words written by each group & their gains

| Group | | N | Minimum | Maximum | Mean | Std. Deviation |
|---|---|---|---|---|---|---|
| Treatment | Total Pretest | 26 | 165 | 457 | 280.58 | 77.99 |
| | Total Posttest | 26 | 219 | 682 | 359.08 | 119.54 |
| | Total Gain | 26 | -67 | 240 | 78.50 | 74.25 |
| Control | Total Pretest | 31 | 171 | 490 | 289.42 | 73.12 |
| | Total Posttest | 31 | 162 | 566 | 304.39 | 84.27 |
| | Total Gain | 31 | -42 | 113 | 14.97 | 30.39 |

Both groups started the course almost at the same level, Mann-Whitney U = 369.50, $z$ = -.54, $p$ = .59, and both groups showed a significant improvement over time, Wilcoxon $z_{(treatment)}$ = -3.92, $p$ < .0005; $z_{(control)}$ = -2.56, $p$ = .01. However, the Mann-Whitney test run between the gain scores of the two groups showed a significant difference, U = 181.50, $z$ = -3.55, $p$ < .0005, with the treatment group being able to outperform the control group.

Maybe the most straightforward measure of grammatical complexity is the frequency of the dependent clauses used by learners as it lends itself well to interpretation and analysis because it is in the form of frequency rather than ratio and is affected by one variable only rather than two. Table 3 presents the descriptive statistics for the two groups across time. As the table shows, both groups could improve in the mean number of dependent clauses they used. While the control group could improve by an average number of one dependent clause over time, the treatment group showed a gain of more than 4.

**Table 3**: The number of dependent clauses used by each group & their gains

| Group | | N | Minimum | Maximum | Mean | Std. Deviation |
|---|---|---|---|---|---|---|
| Treatment | Total Pretest | 26 | 3 | 28 | 13.19 | 6.08 |
| | Total Posttest | 26 | 5 | 39 | 17.77 | 10.04 |
| | Total Gain | 26 | -11 | 27 | 4.58 | 7.93 |
| Control | Total Pretest | 31 | 5 | 27 | 13.32 | 5.35 |
| | Total Posttest | 31 | 6 | 29 | 14.48 | 5.52 |
| | Total Gain | 31 | -4 | 5 | 1.16 | 2.30 |

Based on the results of the Mann-Whitney U test run on the gain scores a trend was observed, U = 288, z = -1.85, $p$ = .06. However, as the results of the Wilcoxon Signed Rank tests showed, a significant difference was observed for both groups over time from the pretest to the posttest $z_{treatment}$ = - 2.63, $p$ = .01, $z_{control}$ = -2.41, $p$ = .02).

However, the pattern of results for the second measure of complexity, the ratio of the number of dependent clauses to the number of clauses, was different. Both groups showed a decline in this measure from pretest to posttest with no significant difference in the gain of the two groups, t (55) = -.44, $p$ = .66. However, while the control group's decline was found statistically significant, t(30) = 3.35, $p$ = .00, this change from pretest to posttest was not statistically significant for the treatment group, t(25) = 1.41, $p$ = .17. This shows that the treatment group undergoing DSS enjoyed a better position in comparison with the control group, suggesting that even if Truscott's claim regarding the negative effect of teacher corrective feedback on the grammatical complexity of their written texts is correct, DSS can compensate for such an adverse effect. Table 4 presents the descriptive statistics for this measure of complexity.

**Table 4**: The two groups' ratio of the dependent clauses to total number of clauses

| Group | | N | Minimum | Maximum | Mean | Std. Deviation |
|---|---|---|---|---|---|---|
| Treatment | Total Pretest | 26 | .18 | .67 | .45 | .12 |
| | Total Posttest | 26 | .18 | .64 | .41 | .12 |
| | Total Gain | 26 | -.37 | .25 | -.04 | .14 |
| Control | Total Pretest | 31 | .24 | .62 | .43 | .10 |
| | Total Posttest | 31 | .22 | .56 | .40 | .09 |
| | Total Gain | 31 | -.12 | .07 | -.03 | .04 |

In the case of the change in learners' accuracy in the written texts, Truscott's main concern, the data were analyzed using the gain score procedure. The independent samples'*t* test run to compare the two groups' gain scores in accuracy was found significant, $t$ (55) = 2.48, $p$ = .02, Eta squared = .10. Moreover, the improvement for the treatment group was statistically significant, t (25) = -2.82, $p$ = .01 with quite a large effect size (Eta squared = .24), while the observed decline in the case of the control group was not found statistically significant, t (30) = 1.14, $p$ = .26. This means that the treatment group had an advantage over the control group. Table 5 summarizes the descriptive statistics for the two groups' accuracy of written texts.

**Table 5:** Descriptive statistics for the two groups' measure of accuracy over time

| Group | | N | Minimum | Maximum | Mean | Std. Deviation |
|---|---|---|---|---|---|---|
| Treatment | Pretest | 26 | .44 | 1.00 | .80 | .14 |
| | Posttest | 26 | .58 | 1.00 | .86 | .13 |
| | Gain | 26 | -.13 | .28 | .06 | .11 |
| Control | Pretest | 31 | .56 | 1.00 | .78 | .09 |
| | Posttest | 31 | .50 | 1.00 | .76 | .15 |
| | Gain | 31 | -.35 | .35 | -.02 | .15 |

# DISCUSSION

According to the results obtained, both groups significantly improved over time in their overall writing proficiency as assessed by the TOEFL iBT holistic scoring rubric, with the treatment group significantly outperforming the control group. While the improvement for the control group was very steady and slow over time, the treatment group's improvement was more eye-catching. The TOEFL iBT writing scoring rubric consists of 5 band

scores. For the purpose of the present study, they were further divided into 3 levels each, with $5^{+}$ (read as five plus) being the highest score and equivalent to 14 in SPSS data entry and $1^{-}$ (read as one minus) being the lowest score and equivalent to zero. While the treatment group improved by almost 3.5 points, the control group improved by 1.5 points only. If the gain scores are converted back to the 5 band score, the treatment group could improve more than one band score while the control group improved only half a band score.

Having examined the rubric, one will notice that band score 3 (equivalent to scores from 6 to 8 in our system of scoring) is the point of departure between proficient writers and non-proficient writers. The features of a good piece of writing in the form of band score descriptors dramatically change from band score 3 to 4. While the band score descriptor for level 3 starts with 'An essay at this level is marked by *one or more* of the following' [emphasis added], the descriptor in band score 4 starts with 'An essay at this level largely accomplishes *all* of the following" [emphasis added]. Both the treatment and control groups started the course of instruction at the same level at band score 3. However, while the control group could not finish the program at a band score higher than what it had started with, the treatment group could pass the point of departure and finish the course in band score 4, even somewhere very close to band score 5. This could show the relative superiority of the intervention to other traditional methods of feedback provision in affecting different features of learners' writing.

Regarding learners' written fluency, the results of the present study indicate that both groups could significantly improve from pretest to posttest with the DSS group outperforming the control group in the number of words written. In the case of the complexity of texts they wrote, while one of the measures showed significant improvement for both groups, the other measure in the form of ratio showed a significant decline for the control group and no significant change in the case of DSS group. This indicates the superiority of DSS in, if not improving the grammatical complexity of learners' written texts, not letting it decline. Finally, while learners receiving corrective feedback alone did not improve in accuracy, the ones receiving corrective feedback plus DSS did improve in accuracy over time.

One point needs to be commented on. It may be objected that Draft-Specific Scoring is not different from a process approach to writing or from the use of portfolios in writing, which is a well-known technique in second language writing. Since the difference between DSS and a process approach to writing has already been explained in Nemati and Azizi (2013), here we suffice with elaborating on the difference between DSS and portfolio writing.

DSS is different from portfolio in many aspects. Here is how Hamp-Lyons (2006, pp. 140-142) defines portfolio:

A portfolio is a *collection* of the writer's work *over a period of time,* usually a semester or school year. The writer, perhaps aided by classmates or the teacher, makes a *selection* from the collected work through a process of *reflection* on what he or she has done… these three elements – *collection, selection,* and *reflection* – are the core of a portfolio [emphasis added].

She also sees the greatest strength and power of portfolio in its potential for a focus on process. To her, that is why it is usually found in process writing classrooms. Hamp-Lyons and Condon (2000, as cited in Weigle, 2002, p. 199) list some of the characteristics of portfolios including:

- An important characteristic of most portfolio programs is *delayed evaluation* [emphasis added].
- Portfolios generally involve *selection* of the pieces to be included in the portfolio, usually by the student with some guidance from the instructor [emphasis added].
- Delayed evaluation and selection offer opportunities for *student-centered control*, in that students can select which pieces best fulfill the established evaluation criteria and can revise them before putting them into their portfolios [emphasis added].

While, in portfolio, the focus is on the process of writing to reach the final product, DSS works with the products to strengthen the processes involved in developing writing proficiency. In portfolio, delayed evaluation is emphasized. On the other hand, immediate teacher evaluation is the corner-stone in DSS. Unlike portfolio assessment, there is no selection in DSS, and instead of collecting students' works over a long period of time, each writing sample is put away after at most three weeks, that is, instead of defining long term objectives, we invest on short-term objectives in DSS in

order to achieve the long-term objective by the end of the course of instruction.

Since there is no delayed evaluation and selection in DSS, there is less chance for student-centered control over the way the class proceeds, but there is a good chance of control over the way their learning and their final score are shaped. That is why unlike portfolio, which is more suitable in process writing approaches, DSS is more appropriate for contexts in which a product approach to writing is practiced. The process approach may offer numerous advantages over the product approach, but as Widdowson (1990) states about the Humanistic approach to language teaching, not everything that is good works well in all contexts and societies. Portfolio needs students who are more autonomous and self-directive so that they can manage their learning and prepare themselves for the final formal evaluation at the end of the semester. However, experience shows that such an approach does not work well in contexts in which students come from an educational background in which all classes are teacher centered and tests and formal evaluation have always been used as tools to push students to study. In such a system, lack of such an evaluation means lack of any need to study. The same seems applicable to our context as well. While working with portfolios takes a great deal of time on the part of the teacher, DSS is also more efficient for both teachers and students because it helps them see the results of the revisions and use of teacher feedback more vividly because in DSS learners have a yardstick to measure their own progress from sample to sample. Finally, the problems discussed before regarding the effect of grading on learners' lack of attendance to teacher feedback cannot be solved through the use of portfolio.

## CONCLUSION AND IMPLICATIONS

DSS allows teachers to "continue their preferred practices while minimizing the negative effect of grading and changing its weakness to strength. It uses grading as a motivating factor which not only does not divert learners' attention from teacher feedback, but it also ensures their attendance to it" (Nemati & Azizi, 2013, p. 141). Applying this technique can resolve the problems grading can cause for language teachers. This way, learners will not throw their writing samples in the wastebasket as soon as they see the grade on them. Instead, they will go through their mistakes to find out the

reason why they had made such mistakes and how they can correct them to improve their scores. This way the mismatch between teachers' beliefs and practices will be resolved. While teachers know that their grading of learners' writing samples distract their attention from their comments, they continue to do so because they feel they have to (Lee, 2009). In addition, teachers will have a profile of learners' scores to easily come up with the final score and satisfy learners' demand for grading as well as the institutional demands for such an evaluation. Keeping such a profile for each student can also help teachers keep track of their learners' improvements over time.

It seems that in case teachers intend to achieve their goals, they need to be aware of the very important role of motivation in learners' attendance to the feedback teachers provide their students with. If not motivated, learners will not pay attention to teachers' comments, and as a result, they will repeat the same mistakes in the following assignments. Therefore, before adopting any method of feedback provision or any type of feedback, they need to think of a way to motivate them to attend to and apply teacher feedback.

In addition, the present study indicates that there could be more intervening variables between teacher feedback and their effect on learners' new pieces of writing. Motivation was one of them. There could be more that need to be looked for. As Bruton (2009) states, teacher feedback must work and when it does not, one should look for what it is that hinders it.

Teachers, material developers, syllabus designers, and policy makers need to be aware of the fact that not everything that sounds good works well in every context. While some methods or ideas may work well in a Western context, they may render to be useless in an Eastern context (Widdowson, 1990). In an Eastern context, Portfolio assessment, which is a very useful and interesting technique, may not work as well as it does in a Western context. In Eastern contexts, grading plays a more significant role than it does in others, and DSS may meet its requirements much more.

Finally, the necessity to revisit Truscott's (1996) claim about the ineffectiveness or harmfulness of corrective feedback seems evident. Corrective feedback works, but it depends on many factors such as the structure being targeted, learners' proficiency level, or their motivation. If the conditions are not appropriate, corrective feedback may fail to work or it

may even show harmful effects. It is a good idea to check the effect of DSS in first language instruction. The problems concerning the effect of corrective feedback equally exist in L1, and DSS can be a good help there too. It also seems to be a good idea to repeat the same study with three groups instead of two; one group receiving corrective feedback only, another group receiving corrective feedback plus DSS, and the other group receiving no feedback at all. This way it is possible to comment on the effectiveness of corrective feedback without DSS as well.

Draft Specific Scoring has an inherent limitation. It works best in contexts in which grading and the scores learners receive play an important role for the learners and the instructional programs. Otherwise, DSS may lose its benefits because it relies on the gains in scores as a result of the revisions learners do. The extent to which this system works in a western context or any context in which grading is not emphasized upon needs to be researched. However, no matter whether DSS works in one context and not in others, something is clear; DSS is a technique and an example to show the effect of intervening variables in the effectiveness of teacher feedback. If DSS does not work in some context, then other ways of motivating learners to attend to teacher feedback should be sought. In any case, it seems plausible to conclude that corrective feedback works provided that learners are motivated enough to attend to teacher feedback and apply it to their writings.

## References

Azizi, M. (2013). *Draft-specific-scoring: A technique to ensure learners' attendance to teacher feedback in L2 writing* (Unpublished doctoral dissertation). University of Tehran, Iran.

Azizi, M., & Nemati, M. (2018). Draft Specific Scoring and Teacher Corrective Feedback: Hearing Learners' Voice. *Journal of Modern Research in English Language Studies, 5*(4), 1-26. doi: 10.30479/jmrels.2019.10708.1340

Bitchener, J., & Ferris, D. R. (2012). *Written corrective feedback in second language acquisition and writing*. New York, NY: Routledge.

Briggs, D. (1980). A study of the influence of handwriting upon grades using examination scripts. *Educational Review*, *32*(2), 185–193.

Bruton, A. (2009). Improving accuracy is not the only reason for writing, and even if it were. *System, 37*(4), 600-613.

Bruton, A. (2010). Another reply to Truscott on error correction: Improved situated designs over statistics. *System, 38*(3), 491-498.

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.

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

Ferris, D. R. (1999). The case for grammar correction in L2 writing classes. A response to Truscott (1996). *Journal of Second Language Writing, 8*(1), 1–10.

Ferris, D. R. (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*(1), 49-62.

Ferris, D. R., Liu, H., Sinha, A., & Senna, M. (2013). Written corrective feedback for individual L2 writers. *Journal of Second Language Writing, 22,* 307-329.

Guenette, D. (2007). Is feedback pedagogically correct? Research design issues in studies of feedback on writing. *Journal of Second Language Writing, 16*(1), 40–53.

Hamp-Lyons, L. (2006). Feedback in portfolio-based writing courses. In K. Hyland & F. Hyland (Eds.), *Feedback in second language writing: Contexts and issues* (pp. 140-161). Cambridge: Cambridge University Press.

Han, Y. (2017). Mediating and being mediated: Learner beliefs and learner engagement with written corrective feedback. *System, 69,* 133-142.

Han, Y. (2019). Written corrective feedback from an ecological perspective: The interaction between the context and individual learners. *System, 80,* 288-303.

Han, Y., & Hyland, F. (2015). Exploring learner engagement with written corrective feedback in a Chinese tertiary EFL classroom. *Journal of Second Language Writing, 30,* 31-44.

Han, Y., & Hyland, F. (2019). Academic emotions in written corrective feedback situations. *Journal of English for Academic Purposes, 38,* 1-13.

Hughes, D. C., Keeling, B., & Tuck, B. F. (1983). Effects of achievement expectations and handwriting quality on scoring essays. *Journal of Educational Measurement, 20*(1), 65–70.

Klein, J. & Taub, D. (2005). The effect of variation in handwriting and print on evaluation of student essays. *Writing Assessment, 10*(2), 134-148.

Lee, I. (2008). Student reactions to teacher feedback in two Hong Kong secondary classrooms. *Journal of Second Language Writing, 17*, 144-146.

Lee, I. (2009). Ten mismatches between teachers' beliefs and written feedback practice. *ELT Journal, 63*, 13–22.

Lee, I. (2014). Feedback in writing: Issues and challenges. *Assessing Writing, 19*, 1–5.

Li, J., & Barnard, R. (2011). Academic tutors' beliefs about and practices of giving feedback on students' written assignments: A New Zealand case study. *Assessing Writing, 16*, 137-148.

Mawlawi-Diab, N. (2015). Effectiveness of written corrective feedback: Does type of error and type of correction matter? *Assessing Writing, 24*, 16–34.

Nemati, M., & Azizi, M. (2013). Grading, no longer an obstacle to learners' attendance to teacher feedback. *Applied Research on English Language, 2*(2), 129-143.

Robb, T., Ross, S., & Shortreed, I. (1986). Salience of feedback on error and its effect on EFL writing quality. *TESOL Quarterly, 20*(1), 83–93.

Russell, M. (2002). *The influence of computer print on rater scores*. Chestnut Hill, MA: Technology and Assessment Study Collaborative. CSTEEP, Boston College.

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

Swain, M. (2005). The output hypothesis: Theory and research. In E. Hinkel (Ed.): *Handbook of research in second language teaching and learning* (pp. 471–83). Mahwah, NJ: Lawrence Erlbaum Associates.

Truscott, J. (1996). The case against grammar correction in L2 writing classes. *Language Learning, 46*(2), 327–369.

Truscott, J. (2004). Evidence and conjecture on the effects of correction: A response to Chandler. *Journal of Second Language Writing 13*(4), 337–343.

Truscott, J. (2007). The effect of error correction on learners' ability to write accurately. *Journal of Second Language Writing, 16*(4), 255–272.

Truscott, J. (2010). Further thoughts on Anthony Bruton's critique of the correction debate. *System, 38*(2), 626-633.

Weigle, S. (2002). *Assessing writing.* Cambridge: Cambridge University Press.

Widdowson, H. G. (1990). *Aspects of language teaching*. Oxford: Oxford University Press.

Wolfe-Quintero, K., Shunji I., & Hae-Young, K. (1998). *Second language development in writing: Measures of fluency, accuracy and complexity*. Honolulu, HI: Second Language Teaching and Curriculum Center, University of Hawai'i at Manoa.

Zhang, Z., & Hyland, K. (2018). Student engagement with teacher and automated feedback on L2 writing. *Assessing Writing, 36*, 90-102.

Zheng, Y., & Yu, S. (2018). Student engagement with teacher written corrective feedback in EFL writing: A case study of Chinese lower-proficiency students. *Assessing Writing, 37*, 13-24.