The Effects of Dynamic Written Corrective Feedback: A 30-Week Study

K. James Hartshorn
Brigham Young University

Norman W. Evans
Brigham Young University

RW

JOURNAL OF RESPONSE TO WRITING

The Effects of Dynamic Written Corrective Feedback: A 30-Week Study

K. James Hartshorn
Brigham Young University

Norman W. Evans
Brigham Young University

This study addresses several challenges in written corrective feedback (WCF) research. First, scholars have expressed concerns that although studies of focused WCF may benefit some classrooms and may help advance second language acquisition theory, they may not represent ecologically valid methods where comprehensive feedback may be more appropriate. Second, many focused WCF studies only report on learner performance within a narrow list of linguistic features, making it impossible for others to determine any secondary benefits or detriments of the treatment. Finally, many research studies of WCF have been of limited duration, making it difficult to identify longer-term effects of various WCF methods. Therefore, this study is an attempt to address these issues by examining the effects of dynamic WCF over a 30-week period. In addition to analyzing linguistic accuracy, this study examined the effects of dynamic WCF on rhetorical appropriateness, fluency, complexity, and vocabulary development over a 30-week period. While improvements in linguistic accuracy were observed for the treatment group when compared to a control group, no other differences were found. Implications for pedagogy and future research are discussed.

Keywords: Dynamic Written Corrective feedback (WCF), L2 writing, comprehensive feedback

Dynamic written corrective feedback (WCF) was developed as an instructional strategy that targets grammatical, lexical, and mechanical errors and was designed to improve second language (L2) writing for those learners whose academic or professional aspirations demand a high level of linguistic accuracy (e.g., Evans, Hartshorn, McCollum, & Wolfersberger, 2010; Hartshorn et al., 2010). Skill acquisition theory (e.g., DeKeyser, 2001, 2007) shapes the theoretical framework for dynamic WCF, particularly in its claim that practice and feedback need to be abundant in order to facilitate greater automatization. Though many who research WCF continue to advocate for "focused feedback," or limiting the number or type of errors to be targeted, some have questioned the ecological validity of such approaches for many classroom contexts (e.g., Storch, 2010), especially since only a few error types have been examined in the literature on focused feedback. Previous studies of dynamic WCF have demonstrated improved accuracy over a 15-week semester when compared to a traditional process writing course (e.g., Evans, Hartshorn, & Strong-Krause, 2011; Hartshorn et al., 2010; Hartshorn & Evans, 2012). Nevertheless, many students need to improve both the linguistic accuracy and the rhetorical appropriateness of what they write. Moreover, there is a need for more longitudinal data if we are to understand the enduring effects of WCF (e.g., Bitchener & Knoch, 2010). Therefore, the aim of this study was to gather data beyond a single semester to determine the longer-term effects of dynamic WCF within an ecologically valid classroom context.

Review of Literature

Though researchers have examined WCF in L2 writing for the better part of a half-century, publications on this topic have proliferated over the past few decades. Some of the central questions have also evolved. For example, the question of whether or not practitioners should provide WCF in L2 writing contexts (e.g., Truscott, 1996, 2007) has largely given way to efforts to identify the most effective ways to provide WCF. Much of the literature has examined the potential benefits of various types of WCF including what has been called direct and indirect and focused or unfocused WCF. Some scholars have tried to identify potential benefits of

direct feedback where corrections are edited into the writing sample. This is in contrast to indirect feedback where the reader marks the location of the errors without providing the correction, so the writers can experience the cognitive benefits of making the correction. Researchers have also differentiated between indirect feedback that is coded where a symbol conveys metalinguistic information about the specific error types and feedback that is uncoded where errors are identified through some type of marking such as circling or underlining (e.g., Ferris & Roberts, 2001; Robb, Ross, & Shortreed, 1986).

Direct Versus Indirect Corrective Feedback

With conflicting results, the specific effects of direct and indirect feedback remain unclear. While some studies suggest that direct WCF may be more beneficial in particular contexts (e.g., Bitchener & Knoch, 2010; Chandler, 2003; Farrokhi & Sattarpour, 2012; Hashemnezhad & Mohammadnejad, 2012; Van Beuningen, DeJong, & Kuikin, 2012), others provide evidence that indirect feedback may be more effective, whether coded (e.g., Ahmadi-Azad, 2014; Erel & Bulet, 2007; Ferris, 2006; Lalande, 1982) or uncoded (e.g., Lu, 2010). Still other studies, however, have observed no differences across various types of direct and indirect feedback (Bitchener & Knoch, 2009a; Ferris & Roberts, 2001; Robb, Ross, & Shortreed, 1986; Semke, 1984), though many such studies have shown the benefits of some form of WCF over control groups not provided with feedback. Additional study is needed if we hope to be able to generalize about the effects of direct and indirect feedback in various contexts.

Focused Versus Comprehensive Corrective Feedback

Another distinction often made in the literature is between what has been called focused and unfocused, or comprehensive, feedback.[1] While focused feedback targets one or a small number of error types, comprehensive feedback targets many or all errors within the writing sample. Most researchers of L2 writing continue to prefer focused feedback over comprehensive feedback because too much feedback can be unmanageable for both the teacher to provide and the learner to process (e.g., Bitchener, 2008; Bitchener & Knoch, 2009a, 2009b; Bitchener, Young, & Cameron, 2005; Ellis, Sheen, Murakami, & Takashima, 2008; Ferris,

2006; Sheen, 2007; Sheen, Wright, & Moldawa, 2009). Nevertheless, other scholars have questioned the ecological validity of focused feedback where a more comprehensive approach may be needed (e.g., Bruton, 2009, 2010; Storch, 2010; Van Beuningen, 2010). Van Beuningen (2010) has argued that comprehensive WCF is more authentic for many classroom contexts and that "the learning potential of comprehensive WCF deserves more attention" (p. 19). Ellis et al. (2008) have also suggested that "the question of the extent to which WCF needs to be focused in order to be effective remains an important one" and determined "if [WCF] is effective when it addresses a number of different errors, it would be advantageous to adopt this approach" (p. 367).

While many scholars continue to recommend focused feedback, some such as Ferris (2010) have raised concerns over the "strict limits on the number of errors" being studied and the "narrowly defined error categories" (p. 192). Most of these studies have limited their focus to the English article system and English past tense verbs. Accordingly, Storch (2010) and Van Beuningen (2010), for example, have expressed apprehension regarding efforts to generalize about the effectiveness of corrective feedback when many of the available studies are based on so few linguistic features. Thus, scholars such as Bitchener (2009) have affirmed the need for researchers to study a wider array of types of feedback.

Some studies have shown comprehensive feedback to be an effective way to facilitate greater accuracy in L2 writing (e.g., Van Beuningen et al., 2012). Other studies provide evidence to suggest that focused WCF may be more effective (e.g., Sheen et al., 2009; Sun, 2013). Still additional studies have found that both focused and comprehensive WCF facilitate improved accuracy without observing a substantive difference between the two types of corrective feedback (CF; Ellis et al., 2008; Hamlin, 2013; Saeb, 2014).

One reason results continue to be inconsistent may be due to a range of confounding variables that are not well controlled. Evans, Hartshorn, McCollum, et al. (2010) hypothesized three categories of variables believed to account for all development in L2 writing accuracy. These include learner variables, situational variables, and methodological variables. Learner variables include L1 background, goals and motivation, L2 proficiency, learning style, and so on. A number of scholars have identified

the role individual learner differences may play in the efficacy of CF (e.g., Bitchener & Ferris, 2012; Evans et al., 2010; Guénette, 2007). Bitchener and Ferris (2012) have noted that "one of the most glaring gaps in the written CF research base to date has been the lack of consideration of individual student differences—in L1, in L2 education, in L1/L2 literacy, in motivation, learning style, personality and so forth" (p. 118).

Among other important learner variables, the proficiency of the student could potentially have a substantial impact on how well the learner can process and apply feedback. Bitchener and Ferris (2012) point out that lower-proficiency learners may become overwhelmed more readily than higher-proficiency learners and that the higher-proficiency learners may have greater metalinguistic knowledge to better process certain types of feedback. As of yet, we have no standardized methods for identifying and measuring many learner differences. While some of these variables may be fairly straightforward, such as L1 and L2 proficiency, others may be much more difficult to systematize, such as motivation, learning preferences, personality, and so forth.

While learner variables may be the most important consideration under ideal circumstances, other variables could also have a tremendous impact on language development. Situational variables could include the teacher and the physical environment as well as prevailing social, political, or economic conditions shaping the teaching and learning context. Methodological variables could include the instructional design, what is taught, and how it is taught. In addition to the results of studies being inconsistent due to differences associated with uncontrolled learner and situational variables, we argue that the method of teaching and learning is also of great importance (e.g., Hartshorn, Evans, & Tuioti, 2014). Even highly intelligent and motivated learners need effective instruction, practice, and feedback in order to maximize their learning.

Dynamic WCF and Skill Acquisition Theory

Dynamic WCF was designed specifically as an instructional strategy to improve the linguistic accuracy of L2 writing and is based on principles from skill acquisition theory as well as from pedagogical practice and observation. Early researchers into the study and theorizing of skill

acquisition include Anderson (1983), who differentiated declarative knowledge (what one knows) from procedural knowledge (what one can do), and McLaughlin, Rossman, and McLeod (1983), who described cognitive processing becoming more automatic and less of a strain on learner attention. The theory predicts that errors will decline as abundant instruction, practice, and feedback increase (DeKeyser, 2001, 2007), thus allowing the learner to move through stages of declarative knowledge and procedural knowledge toward automatization.

Skill acquisition theory is not without limitations that prevent it from becoming a fully viable theory of L2 language development. For example, it does not account for acquisition orders of various linguistic features that do not seem to coincide with language practice. Second, it does not explicitly account for the possibility that some L2 features may not begin with declarative knowledge (Tavakoli, 2013). Despite these limitations, however, skill acquisition theory's emphasis on abundant practice and feedback has been extremely helpful in accounting for important aspects of language development.

Our attempt to operationalize skill acquisition theory is dynamic WCF. With this strategy, writing tasks and feedback are intended to be meaningful, manageable, timely, and constant. In addition to supporting the ideas behind skill acquisition, these principles are consistent with findings from neuroscience. For example, researchers have shown the importance of making instruction, practice, and feedback meaningful and that meaningfulness is the most important criterion the brain uses to determine what material becomes encoded into long-term memory (Devlin, 2010; Schoenfeld, 1988; Sousa, 2010). Other research in neuroscience has shown the need for corrective feedback to be as immediate as possible in order to maximize the benefits of the chemical processes that underlie learning (Willis, 2010).

At the same time, writing practice and feedback must also be manageable in order to ensure that it continues to be meaningful, timely, and constant. Problems with manageability and the limited attentional capacity of learners have led many researchers and practitioners to utilize focused WCF rather than comprehensive feedback. However, we concur with Van Beuningen (2010), who suggested that comprehensive feedback

is more authentic and useful for the classroom, and we agree with Ellis et al. (2008), who suggested that comprehensive feedback would be a better method for feedback if we could learn to use it effectively (see also Bruton 2009, 2010; Storch, 2010).

In order to make comprehensive WCF manageable, we have operationalized these principles such that learners write a 10-minute paragraph on a near-daily basis and receive detailed feedback in the form of coded symbols on all linguistic errors the following class period.

While feedback on broader rhetorical features may not be feasible with such short writing samples, this length seems quite adequate for addressing linguistic accuracy in a manageable way. Because evidence is inconclusive whether direct or indirect feedback is better, we have chosen to use indirect feedback (coded symbols[2] containing metalinguistic information) for two reasons: first, we believe that the learner with adequate linguistic knowledge will benefit from the cognitive engagement needed to correct the errors, and second, the coded symbols make it easy to track the frequency of particular error types for the benefit of the teacher and the students.

Tracking the types of errors students make can inform learners of those linguistic features that are the most challenging for them as individuals. In aggregate, this information can also ensure that classroom instruction is focused on what is needed the most. Students use this feedback to rewrite the paragraph until it is free of errors. Tally sheets (a listing of error frequencies by type for each paragraph), error lists (an ongoing record of all errors along with their surrounding context), and edit logs (a record of how many edits were needed until the writing was deemed "error free") help students to notice the language they produce (for more details about dynamic WCF, see Evans et al., 2010, 2011; Hartshorn & Evans, 2012; Hartshorn et al., 2010).

Studies of Dynamic WCF

Most studies testing the efficacy of dynamic WCF have generated positive results. In Hartshorn et al. (2010), a group of advanced-low to advanced-mid ESL learners in an intensive English program (IEP) who participated in a 15-week course utilizing dynamic WCF made statistically significant improvements in the linguistic accuracy of their

writing, though no significant differences were observed between the treatment and control group for rhetorical competence, writing fluency, and writing complexity. In another publication, analyses demonstrated that no significant differences were found between the treatment group and the control group for use of count and non-count nouns, singular and plural, and verb construction (i.e., subject-verb agreement and verb tense). Nevertheless, statistically significant improvements were observed for determiner accuracy, semantic accuracy,[3] and lexical accuracy (Hartshorn & Evans, 2012).

Additional studies examined learners in a variety of different contexts. For example, Evans et al. (2011) studied matriculated university students in an ESL context. Though the proficiency level of this group of students was somewhat higher than the students in the IEP study, the results were very similar. While dynamic WCF had a large effect on improved linguistic accuracy for the treatment group, no significant differences were observed between the treatment and control groups for fluency and complexity.

Akiyama and Fleshler (2013) also tested the effects of dynamic WCF in a Japanese as a second language context with first-year students. This was a substantial departure from the higher-proficiency levels studied previously. Nevertheless, they observed significantly greater grammatical accuracy from the treatment group in choice of particles and the construction of predicates when compared to a control group. When evaluating dynamic WCF, the treatment group used descriptions such as "helpful," "efficient," "systematic," and "objective" (p. 59). Though comments were predominantly positive, the most prevalent challenge for students was in the use of their error codes (i.e., not understanding what the error was nor how to fix it). This may have been the result of their lower proficiency.

Lee (2009) also conducted a study of dynamic WCF in an intensive program, though there were two differences from the previous IEP study. First, the proficiency of the learners was intermediate-high (i.e., lower than the target proficiency used in the earlier study). Second, in addition to the treatment, both experimental groups participated in traditional process writing classes. The dynamic WCF in Lee's study, however, replaced the grammar class for the treatment group while the control group participated

in a traditional grammar class. In the previous IEP study, dynamic WCF only replaced the process writing class.

This adjustment in the research design was motivated by the belief that this curricular configuration might be more ecologically valid (i.e., students need both rhetorically focused instruction and practice as well as instruction and practice that emphasizes linguistic accuracy). In her study, Lee (2009) found that although the treatment group slightly outperformed the control group for linguistic accuracy in their writing, the difference was not statistically significant. However, similar to the findings from Akiyama and Fleshler (2013), she found that students greatly favored the dynamic WCF over the exercises and activities associated with traditional grammar instruction.

There could be a number of possible reasons for the comparable increases in linguistic accuracy observed for both groups in Lee's (2009) study. First, this could be associated with the fact that the regimen for both groups included both writing practice and a focus on form. It also could be related to the slightly lower proficiency level and that the treatment group may not have been able to process or benefit from the feedback in the same way higher-proficiency students had. Another possibility could be simply that the study was not long enough. Though the higher gains in linguistic accuracy observed for the treatment group were not significant over the course of one semester, one might well ask if the outcome would be different over a longer experimental period.

Research Questions

In addition to the need for more longitudinal data, research in WCF needs to be more complete in what is reported. Many publications studying focused feedback only tell a small part of the story. For example, if a study is limited to English articles or past tense verbs, such studies should not only report on learner performance of these specific features but also provide as much additional detail about other aspects of learner performance as possible. We need to understand whether gains in one aspect of writing come at the cost of another aspect of writing. Some of the most fundamental features of linguistic development that should be provided

to help contextualize accuracy could include rhetorical appropriateness, fluency, complexity, and lexical development (e.g., Skehan, 2009; Skehan & Foster, 2008). With these considerations in mind, we formed the following research questions:

1. What are the longitudinal effects on linguistic accuracy from the combination of a traditional process writing course and either a course emphasizing dynamic WCF or a traditional grammar course?
2. What are the treatment effects on (a) fluency, (b) complexity, and (c) lexical development?

Method

In order to answer these research questions, this study used a mixed model analysis of variance based on a pretest and posttest design after 30 weeks of instruction. This section will briefly address the measures of writing development used in this study, the participants, the procedures, and our reliability estimates.

Measures of Writing Development

For the purposes of this study, linguistic accuracy was defined in terms of the error-free clause ratio as recommended by Evans, Hartshorn, Cox, and Martin de Jel (2014). This was operationalized as the number of error-free clauses divided by the total number of clauses in a sample of writing. Error-free clauses were defined as those without grammatical, lexical, or mechanical errors.

Rhetorical appropriateness was defined with a modified version of the iBT rubric used previously by Hartshorn et al. (2010), which includes six separate levels (0–5). However, preliminary practice with the rubric revealed a potential concern. While the rubric functioned well in the original IEP study, longitudinal data at this proficiency range resulted in an excessive number of writing samples pooling in the middle of the rubric (i.e., ETS Level 3, see Appendix). Therefore, to improve the functional capacity of the rubric to differentiate writing samples, raters were asked to subdivide Level 3 of the rubric into high, mid, and low ratings. Additional practice suggested that this adjustment to eight levels (0–7) allowed the rubric to function better without disrupting the consistency of the raters.

The remaining linguistic features included fluency, complexity, and lexical development. Fluency was defined as the tokens (i.e., total number of words) produced within the 30 minutes allocated for the writing task. Though the treatment used 10-minute paragraphs to ensure the manageability of the feedback, 30-minute essays were used for the pre- and posttests (the same as previous studies of dynamic WCF) because the intent was to see the effects of the treatment on writing tasks that were more rhetorically complex. Complexity was defined in two ways: the first was the mean length of T-unit (MLTU), calculated as the total number of words divided by the total number of T-units in a writing sample; and the second type of complexity was the clause to T-unit ratio, calculated as the total number of clauses divided by the total number of T-units.

Additional measures of lexical development were also considered, including the first and second thousand most frequent words which make up the general service list (GSL; West, 1953), vocabulary from the Academic Word List (AWL; Coxhead, 2000), words not included in the GSL or AWL (off-list words), types (i.e., the number of different words in the text), the type-token ratio, lexical density (i.e., the number of content words over the total number of words), token-family[4] ratio, and the type-family ratio.

These results were easily determined through computational analysis of each writing sample. However, we recognized that misspelled words would automatically be considered "off list" and had the potential to distort the results. Therefore, raters were asked to carefully evaluate each misspelled word and correct the misspelling before running the analysis if they felt highly confident that they knew what word the author intended. If raters were unsure of the intended word, they were asked to delete the word prior to the analysis. While we recognized that this approach could influence the results, we believed that it was the best way to minimize distortion in the lexical analyses.

Participants

The writing of 27 learners was examined in this study including 15 from the treatment group and 12 from the control group. All of these learners were enrolled in the same intensive English program in the United

States. Also, through a battery of placement tests, each student had been placed into the same intermediate proficiency level, approximately 47–63 on the internet-based Test of English as a Foreign Language (TOEFL iBT) or 4 on the International English Language Testing System (IELTS). Because the attrition rate was high, a second cohort of students was added to the control group with the new students beginning the same courses one semester after the original group. Despite these efforts to have equally sized treatment and control groups, after 30 weeks the two control subgroups only produced data for six learners each, making a total of 12 students in all for the control group.

Pretest results were analyzed to help determine whether these groupings would be adequate. The pretest was the same for all of the 27 students. With program teachers and administrators as their audience, students were invited to use narrative and persuasive argumentation to respond to the prompt, *What makes a good teacher? Talk about a teacher you have had that was good and explain why.* Students were given 30 minutes to complete the task. Two t-tests were conducted to help determine the comparability of the writing accuracy from students in these groups. The first test compared the two control subgroups using error-free clause ratios (EFCR). We found no statistically significant difference in EFCR between control subgroup 1 ($n = 6$, $M = .170$, $SD = .108$) and control subgroup 2 ($n = 6$, $M = .142$, $SD = .133$) at the pretest, $t(10) = .408$, $p = .692$. Nor did we find a statistically significant difference between the treatment group ($n = 15$, $M = .215$, $SD = .124$) and the composite control group ($n = 12$, $M = .156$, $SD = 116$) at the pretest, $t(25) = -1.27$, $p = .215$. Because samples were small, these results were interpreted very cautiously. Nevertheless, we proceeded with the assumption that combining the two control subgroups might be acceptable if subsequent analyses supported the comparability of these groups.

The treatment and control groups were similar in many respects. For example, the mean age of students in the treatment group was 25 years and 10 months while the mean age of students in the control group was 24 years and 7 months. The gender and L1 backgrounds of the learners from each group are shown in Table 1.

Table 1

L1 and Gender of Learners in the Experimental Groups

| Native language | Experimental groups | | | | | |
| | Treatment | | | Control | | |
	Male	Female	Total	Male	Female	Total
French	0	1	1	0	0	0
German	0	1	1	0	0	0
Japanese	0	2	2	0	0	0
Korean	1	2	3	1	1	2
Mandarin	0	0	0	1	0	1
Portuguese	1	0	1	0	1	1
Romanian	0	0	0	0	1	1
Russian	1	1	2	0	1	1
Spanish	2	3	5	0	6	6
Totals	5	10	15	2	10	12

Procedures

The treatment and control. Learners in the treatment and control groups participated in four IEP courses four days per week, Monday through Thursday. On Fridays, students in both groups took a number of tests or quizzes relating to their coursework. Four times during each semester, learners in both groups wrote a timed essay like the one used for the pre- and posttests. In three of the four class periods teachers followed the same curriculum for classes emphasizing reading, writing, listening, and speaking. The fourth class period was dedicated to either the treatment or the control (see Table 2). The treatment consisted of linguistic accuracy instruction, writing practice, and the provision of dynamic WCF, as described previously, while the control group consisted of a traditional grammar class.

Table 2

Curricular Regimens for the Experimental Groups

	Experimental groups			
	Control group		Treatment group	
Course emphases	Class time	Homework	Class time	Homework
Reading	4 hrs 20 min	2 hrs	4 hrs 20 min	2 hrs
Writing	4 hrs 20 min	2 hrs	4 hrs 20 min	2 hrs
Listening & speaking	4 hrs 20 min	2 hrs	4 hrs 20 min	2 hrs
Control/treatment	4 hrs 20 min	2 hrs	4 hrs 20 min	2 hrs
Totals	17 hrs 20 min	8 hrs	17 hrs 20 min	8 hrs

The teachers of the grammar classes focused on form, meaning, and use and expended great effort to ensure that explanations were clear and accessible. They also provided students with extensive opportunities to practice through the use of written exercises and various communicative activities. They utilized a popular text that included content such as tense and aspect, adverbial phrases and clauses, passives, intensifiers, complements, comparatives, connectors, relative clauses, hypothetical statements, articles, possessives, and so on. In addition, some activities provided students with the opportunity to integrate the grammar they were learning in various reading, writing, listening, and speaking tasks.

Data elicitation. As described previously, the same pretest was used for all of the students in this study with no significant differences between the treatment and control groups. Nevertheless, in following the IEP's procedure for ensuring test security, prompts were rotated and only half of the control group had the same prompts after 15 weeks and after 30 weeks of instruction. The second half of the control group responded to different prompts as outlined in Table 3. This was a concern because the nature of a prompt may influence the quality of what the learner produces (e.g., Evans et al., 2014; Oliver, 1995; Way, Joiner, & Seaman, 2000). An additional t-test was conducted, which compared the two control groups on error-free clause ratios after 30 weeks. The test showed no significant difference between the two control subgroups, $t(10) = -.085$, $p = .934$. Again, we interpreted these results cautiously due to the small samples.

Table 3

Pretest and Posttest Prompts

Occasion	Treatment & control sub group 1	Control subgroup 2
Pretest prompts	*What makes a good teacher? Talk about a teacher you have had that was good and explain why.*	*What makes a good teacher? Talk about a teacher you have had that was good and explain why.*
Prompts at 30 weeks	*In your opinion, what is the most important characteristic (for example, honesty, intelligence, a sense of humor) that a person can have to be successful in life? Use specific reasons and examples from your experience to explain your answer.*	*In general, people are living longer now. Discuss the causes of this phenomenon. Use specific reasons and details to develop your essay.*

Rating. Two veteran writing teachers with relevant graduate degrees were trained as raters for this study. They were given the task of determining each of the linguistic measures mentioned previously for each writing sample, such as the number of clauses, the number of error-free clauses, the number of T-units and so on. For these tasks, they used an approach of absolute agreement, which required them to be completely unified in their responses for each essay. When occasional discrepancies occurred, they discussed each item until they reached a unanimous decision. The principal researcher also reviewed their work and identified no cases that warranted additional changes. Once the rating was completed, a third rater with minimal instruction and no practice or interaction with the previous raters provided an additional set of rating to establish inter-rater reliability. These data were then used to calculate a mixed model analysis of variance, which analyzed the effects of the treatment by group and time.

Reliability Estimates

Before conducting the planned analyses, we examined the reliability estimates associated with our various measures of writing. Table 4 presents these results. Though we were pleased with most of our reliability estimates, we saw that ratings for the off-list words, rhetorical competence, and accuracy were lower than we had hoped. Nevertheless, we believed that these were adequate to allow us to proceed with our planned statistical analyses.

Table 4

Reliability of Measures

Category	Variable	r
Lexical development	First thousand words	.99
	Second thousand words	.92
	Academic word list	.98
	Off-list words	.73
	Types	.99
	Type/token ratio	.99
	Lexical density	.99
	Token/family ratio	.99
	Type/family ratio	.99
Additional measures of linguistic development	Rhetorical competence	.82
	Fluency (number of tokens)	.99
	Complexity 1 (MLTU)	.98
	Complexity 2 (clauses per t-unit)	.97
	Accuracy (EFCR)	.81

Results

The following results examine accuracy within the context of rhetorical appropriateness, fluency, complexity, and lexical development.

Rhetorical Appropriateness

We begin with rhetorical appropriateness because it provides the needed context to make the other measures of writing meaningful. Table 5 presents the descriptive statistics for the rhetorical appropriateness scores provided by the raters using the rubric found in the Appendix. Learners made substantial progress with their rhetorical appropriateness as a

combined group over the experimental period, $F(1, 25) = 21.316$, $p < .001$, $\eta_p^2 = .460$. Nevertheless, there was no statistically significant difference between the control group and treatment group, $F(1, 25) = 2.031$, $p = .167$.

Table 5

Descriptive Statistics for Rhetorical Appropriateness

	Control ($n = 12$)		Treatment ($n = 15$)		Total ($n = 27$)	
Test	M	SD	M	SD	M	SD
Pre	2.833	1.387	4.367	1.642	3.685	1.694
Post	5.042	1.453	5.533	1.494	5.315	1.469

Fluency

Fluency was defined in this study as the total number of words produced in the 30-minute elicitation period. Table 6 shows the descriptive statistics for the control and treatment groups for the pre- and posttests. As with rhetorical appropriateness, the combined group increased their fluency substantially, $F(1, 25) = 21.518$, $p < .001$, $\eta_p^2 = .463$. However, the differences between the control and treatment groups were negligible, $F(1, 25) = 1.843$, $p = .187$.

Table 6

Descriptive Statistics for Writing Fluency

	Control ($n = 12$)		Treatment ($n = 15$)		Total ($n = 27$)	
Test	M	SD	M	SD	M	SD
Pre	239.67	97.68	361.73	91.14	307.48	111.04
Post	354.92	100.56	424.80	102.50	393.74	105.77

Complexity

Two measures of complexity were used in this study. The first was the MLTU. Table 7 presents descriptive statistics for the control and treatment groups across the pre- and posttests. No significant difference was observed over time for the combined group, $F(1, 25) = 1.796$, $p < .192$. Nor was there a meaningful difference between the control and treatment groups, $F(1, 25) = .014$, $p = .908$. Table 8 illustrates similar results with the

Hartshorn, K. James, and Norman W. Evans. (2015). "The Effects of Dynamic Written Corrective Feedback: A 30-Week Study." *Journal of Response to Writing, 1*(2): 6–34.

second measure of complexity, clauses per T-unit. There was no statistically significant difference over time for the combined group, $F(1, 25) = .155$, $p = .697$. Nor was there a significant difference between the control and treatment groups, $F(1, 25) = 1.775$, $p = .195$.

Table 7

Descriptive Statistics for Writing Complexity 1 (MLTU)

	Control ($n = 12$)		Treatment ($n = 15$)		Total ($n = 27$)	
Test	*M*	*SD*	*M*	*SD*	*M*	*SD*
Pre	8.817	1.489	7.634	1.221	8.160	1.449
Post	9.471	2.229	8.413	3.038	8.883	2.712

Table 8

Descriptive Statistics for Writing Complexity 2 (Clauses per T-unit)

	Control ($n = 12$)		Treatment ($n = 15$)		Total ($n = 27$)	
Test	*M*	*SD*	*M*	*SD*	*M*	*SD*
Pre	1.840	.452	1.683	.382	1.753	.414
Post	1.753	.349	1.803	.540	1.803	.459

Lexical Development

The final variables we consider before examining linguistic accuracy include an array of measures of lexical development. Of the nine variables of interest, eight did not produce a probability of .05 or smaller, suggesting that differences between the control and treatment group were not statistically significant. These included the first thousand most frequent words (K1), vocabulary from the AWL, off-list words, the number of types, the type-token ratio, the lexical density, the token-family ratio, and the type-family ratio as displayed in Table 9. However, there was a statistically significant difference between the control and treatment groups in terms of the relative proportion of the second thousand most frequent words (K2) that were used, $F(1, 25) = 4.44$, $p = .045$, $\eta_p^2 = .151$. This shows that learners in the control group used relatively more words from this category than did those in the treatment group.

Table 9

Summary of Lexical Development

Variable		Control (*n* = 12)		Treatment (*n* = 15)		*df* (1, 25)	
		M	*SD*	*M*	*SD*	*F*	*p*
K1	Pre	.960	.030	.965	.018	.85	.369
	Post	.886	.041	.902	.038		
K2	Pre	.026	.017	.029	.016	4.44	.045
	Post	.059	.032	.036	.015		
AWL	Pre	.021	.019	.015	.012	2.85	.104
	Post	.031	.014	.041	.031		
Off List	Pre	.023	.017	.023	.013	.11	.740
	Post	.028	.012	.026	.016		
Types	Pre	106.00	37.18	139.37	20.05	3.33	.080
	Post	141.92	26.81	154.77	25.73		
Type-Token	Pre	.488	.128	.396	.058	2.60	.119
	Post	.411	.056	.375	.062		
Lexical Density	Pre	.441	.052	.428	.033	.74	.399
	Post	.508	.064	.475	.036		
Token-Family	Pre	2.637	.600	3.236	.560	1.56	.223
	Post	3.048	.448	3.373	.620		
Type-Family	Pre	1.195	.068	1.225	.036	.20	.660
	Post	1.193	.051	1.213	.0387		

Accuracy

The final analysis we examine is linguistic accuracy, which we operationalized using error-free clause ratios. Table 10 provides the descriptive statistics for the control and treatment groups at the pre- and posttest. This analysis revealed a statistically significant difference, showing greater gains in linguistic accuracy for the treatment group when compared to the control group, $F(1, 25) = 8.717$, $p = .007$, $\eta_p^2 = .259$. The interaction effect is illustrated in Figure 2. Since the control group included two subgroups (C1 and C2), mean performance from each subgroup is

Hartshorn, K. James, and Norman W. Evans. (2015). "The Effects of Dynamic Written Corrective Feedback: A 30-Week Study." *Journal of Response to Writing, 1*(2): 6–34.

also plotted in Figure 2, which appears to provide visual confirmation of similarity between control subgroups and the observed differences between the combined control and treatment groups.

This difference is accentuated by additional analyses. While posttest results comparing C1 and C2 were not significantly different, as noted previously, statistically significant differences were observed independently in comparisons between the treatment group and C1 ($M = .161$, $SD = .108$), $F(1, 19) = 8.120$, $p = .01$, $\eta_p^2 = .299$), and between the treatment group and C2 ($M = .166$, $SD = .089$), $F(1, 19) = 8.109$, $p = .01$, $\eta_p^2 = .299$). Though these subgroups are small, they are adequately sized for these tests since they are statistically significant and are accompanied by large effect sizes (see de Winter, 2013). These findings support the general comparability of C1 and C2.

Table 10

Descriptive Statistics for Linguistic Accuracy

Test	Control (n = 12)		Treatment (n = 15)		Total (n = 27)	
	M	SD	M	SD	M	SD
Pre	.156	.116	.215	.124	.189	.122
Post	.164	.094	.362	.158	.274	.165
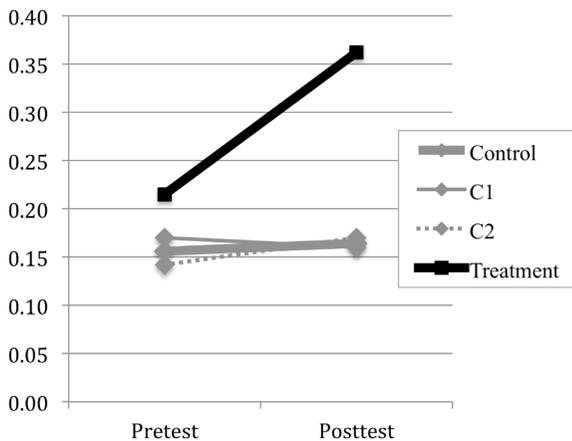
Discussion

Ultimately, this study provides additional support for the extensive practice and feedback advocated by skill acquisition theory. These results show that over a 30-week period, the treatment (a combination of a traditional writing class plus a dynamic WCF class) produced no significant group differences for rhetorical competence, fluency, or complexity when compared to a control (a combination of a traditional writing class plus a traditional grammar class). Nevertheless, the treatment produced statistically significant gains in linguistic accuracy compared to the performance of the control group. The large effect size observed for accuracy in this study ($\eta_p^2 = .259$) seems consistent with previous studies. For example, in the IEP study (Hartshorn et al., 2010) the effect of dynamic WCF on accuracy also was large ($\eta_p^2 = .21$) as was the effect of dynamic WCF on accuracy in the study of matriculated university students ($\eta_p^2 = .16$) though the relative magnitude was somewhat smaller (Evans et al., 2011).

While these studies provide evidence of the possible benefits of dynamic WCF used over time, it may be important to recall that the current study differed from the previous studies in two important ways: first, the dynamic WCF in the current study was used in combination with a traditional writing class (rather than replacing the writing class), and second, the treatment was provided over the course of two semesters rather than only one semester. While one might expect the relative benefit of the treatment in the current study to be much more substantial than those observed in the studies lasting only one semester, the effect size from the current study is only moderately larger than the effect size observed from the previous IEP study. One possible explanation for this could be the inclusion of a separate writing class for both the treatment and control groups, which may have made the overall experience of the two experimental groups more similar and may have benefited both groups in terms of improved accuracy.

While most analyses of the lexis showed no differences, when compared to the treatment group, a significantly greater proportion of the control group's vocabulary came from the second thousand most frequent words as identified by West (1953). Because these analyses were based

on relative proportions within each category, this result would make the most sense if this finding had been accompanied by a smaller proportion for the control group in another category of vocabulary, such as the first thousand most frequent words, the AWL, or the off-list words. Though we see that a relatively smaller proportion of vocabulary from the control group was from the AWL compared to the production of the treatment group, this difference was not statistically significant. The reasons behind this observation remain unclear and additional study is needed.

Pedagogical Implications

This study has shown that dynamic WCF can be used effectively to improve linguistic accuracy as a replacement for a grammar class when accompanied by a traditional process writing class. Though the various components of dynamic WCF were implemented quite consistently in this study and across the previous studies (e.g., Evans et al., 2011; Hartshorn & Evans, 2012; Hartshorn et al., 2010; Lee, 2009), we underscore that the underlying principles of ensuring that writing tasks and feedback are meaningful, manageable, timely, and constant may be more important than strict adherence to any one part of dynamic WCF.

Limitations and Future Research

Several limitations should be considered when evaluating the findings of this study. First is the potential problem of combining two control subgroups into one control group. Though reasonable effort was exerted to ensure a parallel experience and additional analyses suggest the comparability of the two groups, the asynchronous nature of their involvement could have affected the results. Another limitation was the absence of a delayed posttest. Though the decision to forgo a delayed posttest was intentional in order to extend the treatment period to two full semesters, we recognize the benefits a delayed posttest might have provided. Another challenge was the limited functionality of the rhetorical competence rubric. Although we expanded how we used the rubric, which seemed to improve its function, we recognize that this approach may have slightly reduced our reliability. We recommend expanding the rubric before conducting additional longitudinal studies at this proficiency range.

In addition to including a delayed posttest and resolving possible

rubric challenges, we have a number of recommendations for researchers who may replicate or build on this study. For example, if dynamic WCF is used in place of a grammar class as was done here, we recommend giving both groups (experimental and control) the grammar class final exam (if one exists). This would allow researchers to determine whether the group exposed to dynamic WCF would be disadvantaged in the course assessment for not participating in the traditional grammar class. Finally, we recognize that many different contexts exist in which dynamic WCF can be tested. We hope that researchers will choose to build on this research in alternate settings.

Conclusion

The result of this 30-week study showed improvement in the linguistic accuracy of those L2 writers exposed to dynamic WCF while there was no significant difference compared to the control group in terms of rhetorical appropriateness, fluency, or complexity. These findings provide additional evidence that dynamic WCF can be implemented effectively when it accompanies a traditional process writing class. Though we recognize that accuracy in L2 writing is not of equal importance for all learners in all contexts, we believe that it is essential for specific learners in particular contexts. If true, it may become increasingly important in such contexts for learners to receive more meaningful and manageable practice and feedback as they continue to develop their L2 writing.

Notes

1. We will use the term comprehensive feedback to focus on what the feedback is rather than what it is not. More importantly, comprehensive feedback avoids negative connotations associated with the word unfocused (e.g., something lacking purpose or direction).

2. These are the same symbols used in previous studies (i.e., Evans et al., 2011; Hartshorn et al., 2010).

3. This included production choices "which help writers to avoid language that is awkward, unclear, or simply unintelligible" (Hartshorn & Evans, 2012, p. 238).

4. Here family refers to all forms of a word. For example, write, rewrite, and writer would be part of the same family.

References

Ahmadi-Azad, S. (2014). The effect of coded and uncoded written corrective feedback types on Iranian EFL learners' writing accuracy. *Theory and Practice in Language Studies, 4*, 1001–1008. doi:10.4304/tpls.4.5.1001-1008

Akiyama, Y., & Fleshler, M. (2013). Effects of dynamic written corrective feedback on JFL students' homework. In K. Kondo-Brown, Y. Saito-Abbot, S. Satsutani, M. Tsutsui, & A. K. Wehmeyer (Eds.), *New perspectives on Japanese language learning, linguistics, and culture* (pp. 49–68). Honolulu, HI: National Foreign Language Resource Center.

Anderson, J. R. (1983). *The architecture of cognition*. Cambridge, MA: Harvard University Press.

Bitchener, J. (2008). Evidence in support of written corrective feedback. *Journal of Second Language Writing*, 17, 102–118.

Bitchener, J. (2009). Measuring the effectiveness of written corrective feedback: A response to "Overgeneralization from a narrow focus: A response to Bitchener (2008)." *Journal of Second Language Writing*, 18, 276–279.

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

Bitchener, J., & Knoch, U. (2009a). The relative effectiveness of different types of direct written corrective feedback. *System*, 37, 322–329.

Bitchener, J., & Knoch, U. (2009b). The value of a focused approach to written corrective feedback. *ELT Journal*, 63, 204–211.

Bitchener, J., & Knoch, U. (2010). The contribution of written corrective feedback to language development: A ten month investigation. *Applied Linguistics*, 31, 193–214.

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, 191–205.

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

Bruton, A. (2010). Another reply to Truscott on error correction: improved situated designs over statistics. *System*, 38, 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, 267–296.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, 34, 213-238.

DeKeyser, R. (2001). Automaticity and automatization. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 125–151). Cambridge, UK: Cambridge University Press.

DeKeyser, R. (2007). Skill acquisition theory. In B. VanPatten & J. Williams (Eds.), *Theories in second language acquisition* (pp. 97–113). Mahwah, NJ: Lawrence Erlbaum.

Devlin K. (2010). The mathematical brain. In D. A. Sousa (Ed.), *Mind, brain, & education* (pp. 163–178). Bloomington, IN: Solution Tree Press.

de Winter, J. C. F. (2013). Using the student's t-test with extremely small sample sizes. *Practical Assessment, Research & Evaluation*, 18, 1–12.

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, 353–371.

Erel, S., & Bulut, D. (2007). Error treatment in L2 writing: A comparative study of direct and indirect coded feedback in Turkish EFL context. *Journal of Institute of Social Sciences, Erciyes University*, 23, 397–415.

Evans, N., Hartshorn, K. J., & Allen, E. (2010). Written corrective feedback: Practitioner perspectives. *International Journal of English Studies*, 10, 47–77.

Evans, N. W., Hartshorn, K. J., Cox, T., & Martin de Jel, T. (2014). Measuring written linguistic accuracy with weighted clause ratios: A question of validity. *Journal of Second Language Writing*, 24, 33–50.

Evans, N., Hartshorn, K. J., McCollum, R. M., & Wolfersberger, M. (2010). Contextualizing corrective feedback in L2 writing pedagogy. *Language Teaching Research*, 14, 445–463.

Evans, N., Hartshorn, K. J., & Strong-Krause, D. (2011). The efficacy of dynamic written corrective feedback for university-matriculated ESL learners. *System*, 39, 229–239.

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, 49–57.

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

Ferris, D. R. (2010). Second language writing research and written corrective feedback in SLA: Intersections and practical applications. *Studies in Second Language Acquisition, 32*, 181–201.

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

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

Hamlin, R. K. (2013). *Marking time: Is there a differential effect on L2 written accuracy following focused or unfocused written corrective feedback?* (Unpublished master's thesis.) Massey University, Manawatu, New Zealand.

Hartshorn, K. J., & Evans, N. W. (2012). The differential effects of comprehensive corrective feedback on L2 writing accuracy. J*ournal of Linguistics and Language Teaching*, 3, 217–247.

Hartshorn, K. J., Evans, N. W., Merrill, P. F., Sudweeks, R. R., Strong-Krause, D., & Anderson, N. J. (2010). Effects of dynamic corrective feedback on ESL writing accuracy. *TESOL Quarterly*, 44, 84–109.

Hartshorn, K. J., Evans, N. W., & Tuioti, E. A. (2014). Influences on teachers' corrective feedback choices in second language writing. *Writing and Pedagogy*, 6, 251–282.

Hashemnezhad, H., & Mohammadnejad, S. (2012). A case for direct and indirect feedback: The other side of the coin. *English Language Teaching*, 5, 230–239.

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

Lee, S. (2009). The efficacy of dynamic written corrective feedback on intermediate-high learners' writing accuracy. Unpublished master's thesis, Brigham Young University, Provo, Utah, USA.

Lu Y. (2010). *The value of direct and indirect written corrective feedback for intermediate ESL students*. Unpublished master's thesis, Auckland University of Technology, Auckland, New Zealand.

McLaughlin, B., Rossman, T., & McLeod, B. (1983). Second language learning: An information-processing perspective. *Language Learning*, 33, 135–157.

Oliver, E. I. (1995). The writing quality of seventh, ninth, and eleventh graders, and college freshmen: Does rhetorical specification in writing prompts make a difference? *Research in the Teaching of English*, 29, 422–450.

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

Saeb, F. (2014). The effects of focused and unfocused written corrective feedback on the grammatical accuracy of beginner EFL learners. *International Journal of Applied Linguistics & English Literature, 3*, 22–26.

Schoenfeld, A. H. (1988). When good teaching leads to bad results: The disasters of "well-taught" mathematics courses. *Educational Psychologist*, 23, 145–166.

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

Sheen, Y. (2007). The effect of focused written corrective feedback and language aptitude on ESL learners' acquisition of articles. *TESOL Quarterly*, 41, 255–283.

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, 556–569.

Skehan P. (2009). Modeling second language performance: Integrating complexity, accuracy, fluency and lexis. *Applied Linguistics*, 30, 510–532.

Skehan P., & Foster, P. (2008). Complexity, accuracy, fluency and lexis in task-based performance: A meta-analysis of the Ealing research. In

S. Van Daele, A. Housen, F. Kuiken, M. Pierrard, & I. Vedder (Eds.), *Complexity, accuracy, and fluency in second language use, learning, and teaching* (pp. 263-284). Brussels, Belgium: University of Brussels Press.

Sousa, D. A. (2010). How science met pedagogy. In D.A. Sousa (Ed.), *Mind, brain, & education: Neuroscience implications for the classroom* (pp. 9–26). Bloomington, IN: Solution Tree Press.

Storch, N. (2010). Critical feedback on written corrective feedback research. *International Journal of English Studies*, 10, 29–46.

Sun, S. H. (2013). *Written corrective feedback: Effects of focused and unfocused grammar correction on the case acquisition in L2 German.* (Unpublished doctoral dissertation.) University of Kansas, Lawrence, Kansas.

Tavakoli, H. (2013). *A dictionary of language acquisition: A comprehensive overview of key terms in first and second language acquisition*. Tehran: Rahnama.

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

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

Van Beuningen, C. G. (2010). Corrective feedback in L2 writing: Theoretical perspectives, empirical insights, and future directions. *International Journal of English Studies, 10*, 1–27.

Van Beuningen, C. G., De Jong, N. H., & Kuiken, F. (2012), Evidence on the Effectiveness of Comprehensive Error Correction in Second Language Writing. *Language Learning*, 62, 1–41.

Way, P. D., Joiner, E. G., & Seaman, M. A. (2000). Writing in the secondary foreign language classroom: The effects of prompts and tasks on novice learners of French. *The Modern Language Journal, 84*, 171–184.

West, M. (1953). *A General Service List of English Words*. London: Longman, Green.

Willis, J. (2010). The current impact of neuroscience on teaching and learning. In D. A. Sousa (Ed.), *Mind, brain, & education: Neuroscience implications for the classroom* (pp. 45–65). Bloomington, IN: Solution Tree Press.

Appendix

Adjustments Made to the Rhetorical Competence Rubric

Writing Rubric Adapted from the iBT TOEFL Test	
ETS Level	**Description**
𝟝 (X) → 7	**The essay accomplishes the following:** • effectively addresses the topic and task • is well organized and well developed, using clearly appropriate explanations, examples, support or details • displays unity, progression, and coherence
𝟜 (X) → 6	**The essay accomplishes the following:** • addresses the topic and task well, though some points may not be fully elaborated • is generally well organized and well developed, using appropriate and sufficient explanations, examples or details • displays unity, progression, and coherence, though it may contain redundancy, digression, or unclear connections
𝟛 (X) → 5, 4, 3	**The essay is marked by one or more of the following:** • addresses the topic and task using somewhat developed explanations, example or details • displays unity, progression, and coherence, though connection of ideas may be occasionally obscured
2	**The essay may reveal one or more of the following:** • limited development in response to the topic and task • inadequate organization or connection of ideas • inappropriate or insufficient examples or details to support or illustrate generalizations in response to the task
1	**The essay is seriously flawed by one or more of the following:** • serious disorganization or underdevelopment • irrelevant specifics or questionable responsiveness to the task • little or no detail
0	An essay at this level merely copies words from the topic, rejects the topic, is otherwise unconnected to the topic, or is blank.

Hartshorn, K. James, and Norman W. Evans. (2015). "The Effects of Dynamic Written Corrective Feedback: A 30-Week Study." *Journal of Response to Writing, 1*(2): 6–34.