Comprehensive corrective feedback in foreign language writing: The response of individual error categories

Marisela Bonilla López[1,2], Elke Van Steendam[1], Dirk Speelman[1] & Kris Buyse[1,3]

[1]KU Leuven, Leuven | Belgium

[2]Universidad de Costa Rica, San José, San Pedro | Costa Rica

[3]Nebrija University, Madrid | Spain

Abstract: While the literature on the effect of comprehensive corrective feedback (CF) on overall accuracy is abundant, the body of work employing such a scope to explore error treatability is not, especially when it comes to blended (cf. Ferris, 2010) design studies. Consequently, this investigation extends the analyses from the data set of Bonilla et al. (2018) to report on individual linguistic features. Specifically, to address crucial amenability-related questions in need of perusal, the present blended design study explores the effect of two types of comprehensive CF (with direct correction and metalinguistic codes) on the treatability of separate grammatical and non-grammatical structures. To this end, a group of EFL learners (N = 139) were required to do editing that involved error-correction, deferred (on a draft), and focused on language as well as to produce two independent essays (in an immediate and a delayed posttest). Main results from logistic regression (to test the effect in revised essays) and mixed-effect models (to test the effect on independent essays) render seven variables that can explain correctability differences: out of those, three have also explained overall accuracy gains (cf. Bonilla et al., 2018), one has not been identified thus far, and three consolidate themselves as relevant factors under other conditions as well. Theoretical and pedagogical implications are discussed.

Keywords: comprehensive corrective feedback, direct corrections, second language teaching, grammatical errors, non-grammatical errors, metalinguistic codes

journal of WRITING RESEARCH

Bonilla López, M., Steendam, E. V., Speelman, D., & Buyse, K. (2021). Comprehensive corrective feedback in foreign language writing: The response of individual error categories. *Journal of Writing Research, 13*(1), 31-70. https://doi.org/10.17239/jowr-2021.13.01.02

Contact: Marisela Bonilla López, Universidad de Costa Rica / Escuela de Lenguas Modernas, Ciudad universitaria Rodrigo Facio Brenes, San José, San Pedro| Costa Rica – marisela.bonilla@ucr.ac.cr - ORCID: 0000-0002-1194-7721

1. Introduction

Empirical interest in written corrective feedback (CF) (also known as *error correction*) dates as far back as the late 1960's (e.g., Stiff, 1967). As such, not only has it been defined as "feedback on forms with a view to advancing the language learning of the writer and thus contributing to text quality" (Murphy & Roca de Larios, 2010, p. viii), but it has also come a long way. Specifically, because error correction has been (and still is) a ubiquitous practice in second language (L2) (Ferris, 2010) and foreign language (FL) classrooms (Pawlak, 2014), it was assumed to be effective. This means that early studies did not question its effectiveness but tried to explore to what extent different feedback strategies were superior to others (Bitchener & Ferris, 2012). They did so by looking into differential effects from an initial text to its revised version, earning them the label *revision studies* (also *L2 writing* studies in Ferris, 2010; or *feedback for accuracy* studies in Manchón, 2011). However, by questioning these studies because of their lack of theoretical relevance (cf. Truscott, 1996) or their design and execution shortcomings (cf. Guénette, 2007), a new strand of L2 acquisition feedback studies emerged (also *feedback for acquisition* studies in Manchón, 2011). This implies that the interest was no longer on the efficacy of a given feedback strategy to bring about accuracy during text revision but on the extent to which written CF could lead to L2 learning (measured as sustained accuracy gains from pre-test to [immediate or delayed] posttest).

The resulting line of research has given L2 (writing) practitioners and researchers enough evidence on the potential of written CF to develop learners' interlanguage (cf. Bonilla, Van Steendam, Speelman, & Buyse, 2018; van Beuningen et al., 2012) and foster L2 acquisition processes (cf. Storch & Wigglesworth, 2010). In fact, from a cognitive/interactionist perspective (e.g., Long, 1996; Schmidt, 1995, Swain, 1985, 1995), findings that L2 learners can process written CF successfully (as evidenced in short- and long-term gains) mean that they are able to complete "the stages in the development of L2, from the initial written CF input stage to the implicit, automatized output stage" (Bitchener & Storch, 2016, p. 2). Hence, despite being originally intended for oral production, researchers acknowledge that the cognitive/interactionist perspective is the one that has the most room to explain the likely effects of written CF and its L2 learning potential (Bitchener, 2012a; Bitchener & Ferris, 2012; Polio, 2012). For example, applied to L2 writing, Bitchener (2012a) explains that such a standpoint (1) acknowledges the crucial role that input (e.g., negative evidence such as written CF), noticing (e.g., studying the feedback), and output (e.g., a revised or a new text) play in L2 acquisition processes; (2) is clear in that something more than mere L2 exposure is needed for L2 development, hence, the relevance of 'pushed' output for producing modified output; and (3) stresses the importance of attention in facilitating L2 learning (e.g., Schmidt, 1990, 2001).

With this in mind, a strong case for the provision of written CF in the L2 (writing) class in general and text revision in particular is evident.

Even so, the literature on error correction is far from providing conclusive answers in many respects. For example, because L2-acquisition feedback studies tend not to have a revision component in their design, there has been a call to reconcile L2 writing and L2 acquisition feedback research strands (e.g., Ferris, 2010; Sheen, 2010a). The rationale behind it is the need to yield answers that can be theoretically and pedagogically relevant to both lines of inquiry. That is, a pretest-posttest-delayed posttest design looks into what ultimately matters from an L2 acquisition standpoint (i.e., L2 development), but it is still necessary to address a measure that is relevant for L2 composition teachers (i.e., accuracy). For this reason, Sheen (2010) states that "inquiry into written CF within the [L2 acquisition] research paradigm can be seen as relevant to L2 writing pedagogy, given that one of the aims of such pedagogy is to improve students' written … accuracy" (p. 211). Along the same lines, Ferris (2010) highlights that incorporating the revision component is called for. After all, revision is an integral part of the writing process (Ferris, 2004; Fitzgerald, 1987; Fitzgerald & Markham, 1987). Hence, to Ferris (2010), a pretest-delayed-posttest design that also includes revision—which the author labels as "blended" (p. 195)—bridges the gap between two lines of inquiry that have always looked into the same phenomenon but with two differing starting points. Interestingly, those differences are reflected not only in L2 researchers' empirical interests but also in L2 (writing) teachers' reasons to correct learners' written errors. As Bitchener (2012b) explains, "[w]hile composition teachers may be more likely to do this so that their learners can edit their writing and produce error-free revisions, language learning teachers may do so in order to help their learners acquire specific target-like forms and structures, demonstrated in the writing of new texts" (p. 855). Still at this point, it would be unreasonable to believe that L2 writing instructors could not be concerned about whether students are able to write without errors over time. The truth is that "language … matters" (Ferris & Eckstein, 2020, p. 322), and both L2 writing and L2 acquisition practitioners may as well be interested in two outcomes: the feedback effect on accurate language use both in the short- and long-term.

Despite the aforementioned, the incipient body of *blended* feedback studies has exclusively focused on overall accuracy (for a review, see Bonilla, Van Steendam, & Buyse, 2017). Out of them, only Van Beuningen et. al (2012) and Bonilla et al. (2018) have examined two types of overall accuracy: grammatical and non-grammatical. For example, Van Beuningen et al. did so with a sample of secondary school Dutch learners who corrected errors with either direct corrections or codes. Their results showed that learners' grammatical accuracy benefited more from direct corrections and non-grammatical accuracy from codes. Then, in a large-scale study with a research design that included not only a baseline comparison with

different feedback scopes (i.e., groups correcting grammatical errors only or correcting both grammatical and non-grammatical ones) but also an interest on learner-related variables (i.e., cognitive load and attitudes), Bonilla et al. found that English as a foreign language (EFL) learners were more grammatically accurate when their attention was drawn on grammar issues only, that in the long run direct corrections were superior to codes for either type of overall accuracy, and that a higher degree of feedback explicitness both imposed a lower self-reported cognitive load and rendered more feedback comprehensibility. Conflicting findings aside (which may be attributed to design differences), these two studies greatly contribute to widening current knowledge on written CF. However, blended design studies have not explored error treatability yet—typically used to refer to the degree of correctability of an individual error category (e.g., Ferris, 1999; Truscott, 2001).

What is more, the state of the literature thus far reveals that available evidence on the responsiveness of individual linguistic categories emerges mainly from studies that have been conducted in English as a second language (ESL) settings and that are restricted to a handful of targeted features (e.g., Bitchener, 2008). The problem then lies in that previous research findings may not be generalizable to other instructional settings such as the FL class, where error correction tends to address multiple types of errors (i.e., comprehensive CF) (Ellis et al., 2008) and employ feedback strategies that have received little attention in studies of this type (e.g., direct corrections and metalinguistic codes). Interestingly, it has already been pointed out that FL revisers tend to place a high emphasis on editing (low-order concerns such as grammar and stylistics) (cf. Van Steendam, Rijlaarsdam, Sercu, & Van den Bergh, 2010). Thus, there is a sound need to both conduct written CF research that conforms with feedback practices of FL classrooms and invest efforts to shed some light into the variables that may influence error treatability when editing activities are involved.

Consequently, with much-needed research efforts on error correctability in mind, the present study extends the analyses from the data set of Bonilla et al. (2018) to report on individual linguistic features. Specifically, to address crucial amenability-related questions in need of perusal, the present blended design study explores the effect of two types of comprehensive CF (with direct correction and metalinguistic codes) on the treatability of separate grammatical and non-grammatical structures. To this end, a group of EFL learners were required to do editing that involved error-correction, deferred (on a draft), and focused on language (for a complete taxonomy of revision, see Stevenson et al., 2006) as well as to produce two independent essays (in an immediate and a delayed posttest).

2. Empirical background

Thus far, Ferris's (1999) treatable/untreatable dichotomy has been the go-to work to discuss and interpret issues in error treatability (e.g., Bitchener et al., 2005; Ferris, 2010; Ferris & Roberts, 2001). The reason behind this is that according to Ferris, part of the treatability of an error hinges on the existence of a straightforward rule and whether or not it is readily available for learners to consult: if it is, the error is rule-based and likely to be responsive to written CF (i.e., treatable); if it is not, it is non-rule based and untreatable (also see, Ferris, 2006). Nonetheless, this distinction has been recently challenged on the grounds that it "is an ad hoc one with no theoretical basis" (Shintani et al., 2014, p. 7). Despite this criticism, previous empirical efforts have been able to cautiously unravel part of the complex interplay involved in the extent to which a given separate error category responds (or not) to written CF. As a result, three potentially influential variables have been gleaned: feedback type, error type, and feedback scope.

2.1 Feedback type

The specific written CF treatment that learners receive has been pointed out as an explanatory factor for some of the accuracy results obtained to this day (e.g., Bitchener, 2008; Bitchener & Knoch, 2010a; Ferris, 2006). For instance, Ferris (2006) targeted 15 error categories and intended to implement a coding system for supplying feedback to 92 ESL composition students. The author reported high percentages in learners' successful corrections of most of the targeted features. Specifically, learners were the most successful at editing errors in categories such as run-on sentences (87.3%), spelling (85.4%), and singular-plural (84.9%). Nevertheless, the results pertaining to long-term accuracy showed that there was no improvement (from essay 1 to 4) in errors that had rendered high correction percentages (e.g., subject-verb agreement errors) whereas significant progress was observed in those errors that had yielded lower percentages of correction success (i.e., verb-related errors). The fact that the teacher participants employed differing feedback strategies led Ferris (2006) to conclude that such results may have been due to subject-verb agreement errors being corrected mostly with direct corrections and verb errors mostly with metalinguistic codes. The author suggested that the explicit provision of the direct corrections may prove effective enough in the short-term but that the problem-solving nature of the metalinguistic codes may ultimately be more beneficial in the long run. However, one issue with Ferris's (2006) study is that, as the author acknowledged, the singled-out effect of a specific treatment (such as directions and metalinguistic codes) cannot be obtained because the feedback strategies were not systematically implemented.

Studies with a pretest-posttest-delayed posttest design have also yielded evidence of the role that feedback type could play in error treatability (e.g., Bitchener, 2008; Bitchener & Knoch, 2010a). Most of such studies have examined ESL

(university) learners, targeted a narrow number of error categories, and investigated the feedback effect by supplementing feedback strategies (i.e., a combination of strategies). To illustrate, the accuracy with which ESL learners improve their use of the English article has been predominantly examined with direct corrections with or without (oral or written) metalinguistic explanation (e.g., Bitchener & Knoch, 2008, 2010b; Sheen, 2007). In general, results have shown that the English article is likely to be responsive to written CF with supplemented direct corrections (e.g., Bitchener, 2008; Bitchener & Knoch, 2008) although when treated with a single-feedback variable (i.e., unsupplemented) such as underlining, learners' accuracy improvement of the article usage may not occur (e.g., Bitchener & Knoch, 2010a).

In this respect, the degree to which a given treatment affords learners with opportunities to notice the difference between what is expected and what they produce matters. It goes without saying that being given the chance to notice what exactly is wrong with one's output and discussing it (e.g., Bitchener, Young, & Cameron, 2005) or understanding it through an explanation (e.g., Sheen, 2007) may be more likely to enhance noticing processes than underlining (e.g., Bitchener & Knoch, 2010a), which is an indirect feedback strategy that merely locates the error. The relevance of such noticing lies then in its role: it is "crucial to uptake and long-term acquisition" (Bitchener et al., 2005, p. 201). However, one important aspect to take into consideration is how applicable previous findings are to FL writing contexts where such a degree of supplementation of strategies (i.e., grouping variables) is unrealistic because of large class sizes and a heavy workload. The truth of the matter is that out of the studies that have looked into error treatability, most have examined grouping variables (e.g., Bitchener & Knoch, 2009b; Diab, 2015). A few studies have investigated single-feedback variables (e.g., Sheen, Wright, & Moldawa, 2009; Shintani & Ellis, 2013), out of which none has addressed the differential effect of two common feedback strategies in FL writing classes: direct corrections and metalinguistic codes. Thus, the present study constitutes a start in that direction.

2.2 Error type and complexity

Researchers agree that the efficacy of a given treatment is largely influenced by the type of error (e.g., Bitchener, 2012; Shintani & Ellis, 2013) and not solely by feedback type. With this in mind, one of the most common ways to discuss error treatability has been by drawing on the treatable/untreatable dichotomy proposed by Ferris (1999). For instance, in a data reanalysis, Ferris and Roberts (2001) grouped the targeted error categories into treatable (i.e., articles, noun, and verbs) and untreatable ones (i.e., sentence structure and word form). The authors concluded that because comparisons of success ratios reached significance with word choice but not with sentence structure, "some 'untreatable' errors may be more so than

others—specifically, complex sentence structure problems versus single word errors" (p. 173). Also, in discussing their results, Bitchener et al. (2005) explained that their ESL learners were able to significantly improve their accurate use of past tense and definite article because these features were "determined by sets of rules" (p. 201). The authors went on to add that the opposite was true for prepositions, which learners could not improve over time because this target was more idiosyncratic.

Another study that refers to error type to explain differences in accuracy performance is Diab (2015). The author reported differing patterns of response for ESL learners' pronoun agreement and lexical errors and, as Bitchener et al. (2005), Diab (2015) partly attributed such results to the rule-based nature of the targeted structures. The author explained, for instance, that in the case of pronoun agreement errors, a group receiving direct corrections plus metalinguistic codes was able to significantly improve at the immediate posttest because these errors were rule-based, implying that learners were able to retrieve in first drafts the rule they had had the chance to practice before. Conversely, according to Diab (2015), the rule-based notion did not apply to lexical errors, which explained why the same pattern of significance was not obtained in this type of errors. From this perspective, the extent to which learners can internalize written CF could hinge on whether there is a (straightforward) rule to consult or not. Nonetheless, the existence of a rule (or lack thereof) to be accessed readily (i.e, the treatable/untreatable dichotomy) has already been criticized for lacking a theoretical basis (cf. Shintani et al., 2014), making error complexity a more likely factor to explain differences in error treatability (e.g., Shintani et al., 2014).

For example, previous research evidence has shown that learners can improve their accurate use of particular structures under certain conditions even up to two months (e.g., Diab, 2015) and that some error categories have shown to be better candidates for correction than others. Among these are articles (Bitchener et al., 2005; Nassaji, 2011), pronouns (Diab, 2015), and nouns (Ferris & Roberts, 2001). Errors in prepositions (e.g., Bitchener et al., 2005; Nassaji, 2011) and lexical (e.g., Diab, 2015) categories, on the other hand, could be harder to treat based on the evidence thus far. It should be noted, however, that the majority of studies have focused on a narrow number of grammatical issues (e.g., Bitchener & Knoch, 2009b; Ellis et al., 2008; Ferris & Roberts, 2001; Shintani & Ellis, 2015). This means that a larger variety of individual grammatical problems (e.g., word form, fragment, sentence structure) has been addressed to a lesser degree and that separate non-grammatical issues remain unexplored (e.g., spelling, punctuation, and capitalization). Therefore, the literature could benefit from a study that investigates individual error categories that have not received enough empirical attention to this day.

2.3 Feedback scope

By comparing two differing feedback scopes, a few studies have suggested that the amount of written CF that learners receive could have some influence on the accuracy effect of separate errors (e.g., Sheen, Wright, & Moldawa, 2009). Such a claim has emerged from studies which have compared highly selective CF with mid-selective CF (e.g., Ellis et al., 2008; Farrokhi & Sattarpour, 2012; Sheen, Wright, & Moldawa, 2009) or highly selective CF with comprehensive CF (e.g., Frear & Chiu, 2015). Specifically, bearing in mind that the likely effect of a given scope is thought to be influenced by learners' attentional capacity and processing ability (cf. Bitchener, 2012), some researchers have sought to determine what scope is more effective to treat articles (e.g., Ellis et al., 2008; Farrokhi & Sattarpour, 2012; Sheen et al., 2009) or past tense verbs (e.g., Frear & Chiu, 2015). In a nutshell, findings have been mixed. For instance, Ellis et al. (2008) reported that a highly selective treatment and a mid-selective one were equally useful to enhance EFL learners' accuracy improvement of article usage: both scopes significantly outperformed the control group and no significant differences were observed between them.

Nonetheless, Farrokhi and Sattarpour (2012) and Sheen et al. (2009) found otherwise in their studies with EFL and ESL participants, respectively. Their evidence suggested that a highly selective treatment was more advantageous to improve learners' accuracy of article use than a mid-selective approach, being the main reason that the former was less overburdening than the latter, according to the authors. Then, with a different comparison in their design, Frear and Chiu (2015) provided evidence that both selective CF and comprehensive CF can bring about accuracy improvement in the use of past tense verbs but that only a narrow approach to errors can enable ESL learners to significantly improve overall accuracy.

At this point, it is evident that the empirical interest on error treatability has almost exclusively focused on a narrow number of targeted features and that evidence from a broader feedback scope—common in FL writing settings—is insufficient. This void has already been pointed out (cf. Bonilla et al., 2017; van Beuningen et al., 2012), yet no study to the best of our knowledge has delved into the exact effect of comprehensive CF on learners' short- and long-term control of individual features.

3. The current study

The evidence in Bonilla et al. (2018) indicated that although a focus on grammar issues only may be more beneficial for enhancing grammatical accuracy, attention to a larger array of errors does not prevent EFL learners from being both grammatically and non-grammatically accurate. The same study also found that in the long term, direct corrections may be more advantageous than codes to bring about overall grammatical and non-grammatical accuracy. Nonetheless, there is no certainty that a broad feedback focus would render the same results when

examining individual error categories under the same circumstances. For this reason, considering that no blended design study to this day has explored the amenability to correction of separate linguistic features, a valuable addition to the literature would be an extension of the work in Bonilla et al. (2018) to individual error types. Doing so could contribute to gaining much-needed insight into error treatability and the full potential of feedback practices that are more representative of FL contexts.

Specifically, the present study aims to generate theoretical and practical knowledge for L2 acquisition researchers and L2 writing practitioners alike in four main ways: (1) by delving into the value of comprehensive CF as a revision and a learning tool (i.e., a blended design), (2) by targeting separate error categories that have not been examined to date (e.g., non-grammatical features), (3) by exploring error amenability to feedback with two common feedback strategies in FL classes (i.e., unsupplemented direct corrections and metalinguistic codes), and (4) by examining a largely overlooked learner type in error treatability studies (i.e., L2 learners in non-dominant English settings).

3.1 Research questions

The objective of this study was to examine the effect of two types of comprehensive CF (with direct correction and metalinguistic codes) on the treatability of separate grammatical and non-grammatical structures during revision and on later independently written essays. To this end, the students wrote an initial essay, received CF and studied it, and then revised it without continued access to the CF. This was done to answer the first research question (RQ1): What is the effect of comprehensive CF (provided with direct corrections or metalinguistic codes) on the accuracy with which learners correct separate grammatical and non-grammatical errors during essay revision? In addition, students wrote a second essay and received CF but did not revise it. Instead, they wrote an independent essay immediately after that and then another one four weeks later. Doing so would contribute to answering the second research question (RQ2): What is the effect of comprehensive CF (provided with direct corrections or metalinguistic codes) on the accuracy with which learners use separate grammatical and non-grammatical errors on independent essays?

3.2 Expected outcomes

After Truscott's (1996) assertion that morphological, syntactic, and lexical errors represent different domains of knowledge, more researchers acknowledge such a possibility and include it in their discussion on error amenability to written CF (e.g., Bitchener et al., 2005; Ferris, 2010; Ferris & Roberts, 2001). Indeed, the literature has numerous references to some errors being discrete and rule-governed (i.e., rule-based) and others being complex, idiosyncratic, or semantically based (e.g., Diab,

2015; Frear & Chiu, 2015; Guénette, 2012; Kurzer, 2017; Lee, 2013). Such references are informed by Ferris's (1999) dichotomy, which distinguished between treatable errors—those that "occur in a patterned, rule-governed way"—and non-treatable ones—more complex idiosyncratic problems in which no readily available "handbook or set of rules" (p. 6) can be consulted to correct them. Since its implementation in different feedback studies (e.g., Diab, 2015; Ferris, 2006; Ferris & Roberts, 2001), the treatable/untreatable categorization has been a valuable addition to the literature because it has served to claim that the correctability of an error may determine the feedback explicitness needed to correct it (Brown, 2012; Ferris, 1999; Ferris & Roberts, 2001; Guénette, 2012; Hyland, 2003). For example, Brown (2012) explains that more treatable errors in which "students can reference straightforward rules to self-correct" would benefit more from an indirect treatment whereas those "untreatable idiosyncratic errors [which] require students to use acquired knowledge to make corrections" would be more amenable to a direct approach (Brown, 2012, p. 863). However, what exactly constitutes a treatable and an untreatable error remains unclear to this day (Bitchener, 2012a; Lee, 2013).

In fact, to formulate hypotheses, the state of the field is not conclusive as to what characteristics make some errors more complex and in turn more amenable to written CF than others—hence the exploratory nature of this study. Consequently, due to previous criticism to the treatable/untreatable dichotomy (e.g., Shintani et al., 2014), the rule-based nature of an error (or lack thereof), as explained in Ferris (1999) and Brown (2012), was not borne in mind in our predictions. Instead, we drew on the study by van Beuningen et al (2012), which found that overall grammatical accuracy may benefit more from direct corrections whereas overall non-grammatical accuracy may be more responsive to metalinguistic codes. In the light of these findings and applied to individual error categories, two possible outcomes were expected:

a. It seemed plausible that separate grammatical errors such as articles, subject-verb agreement, prepositions, word form, verb, subject deletion, subject repetition, sentence structure, pronouns, and fragments would become good targets with direct corrections.

b. Along these lines, separate non-grammatical errors such as capitalization, spelling, and punctuation were expected to be more amenable to written CF with metalinguistic codes.

4. Methods

4.1 Participants and setting

All student writers in this study ($N = 139$) were enrolled in the School of Modern Languages of the *Universidad de Costa Rica*. They were either English or English Teaching majors (i.e., EFL learners). This entails that all participants were pursuing a

career as English professionals. The course where the participants (53 male and 86 female, mean age = 21, *SD* = 4.11, age range = 18-38) were enrolled was an intensive, introductory first-year English course that placed emphasis on the four macro skills (e.g., speaking, listening, reading, and writing). This means that the participants were novice writers in academic contexts. Thus, learners met four days a week (three hours/three days; four hours/one day). As far as the writing component of the course is concerned, one particular characteristic of these EFL students was that they were learning how to write in English and practicing English through writing—a characterization that fits Manchón's (2011) distinction between the learning-to-write and writing-to-learn language learning dimension.

The participants' proficiency level was lower-intermediate (*M* = 2.27, *SD* = .79), and their native language was Spanish. They were randomly assigned to four experimental conditions and a control group. More specifically, students were randomly assigned to different classes, and within classes, they were randomly assigned to conditions. In the four experimental conditions, participants received feedback on grammatical errors with direct corrections (DG, *n* = 29), grammatical errors with metalinguistic codes (MG, *n* = 28), grammatical and non-grammatical errors with direct corrections (DGN, *n* = 27), or grammatical and non-grammatical errors with metalinguistic codes (MGD, *n* = 28). In the control condition, the participants did not receive written CF but self-corrected their errors to the best of their abilities (C, *n* = 27).

4.2 Feedback scope and strategies

The number of targeted features as well as the feedback strategies employed in this study had to conform to the feedback practices of the instructional context surrounding the treatment. Consequently, the feedback scope was comprehensive, which has been defined as corrections on a wide range of errors (Bitchener & Ferris, 2012; Ellis et al., 2008). Specifically, the targeted error categories were grammatical (i.e., preposition, article, subject-verb agreement, word form, verbs, subject deletion, subject repetition, sentence structure, pronouns, sentence fragment) and non-grammatical (i.e., capitalization, spelling, and punctuation). This means that the error categories dealt with issues on morphology, syntax, and mechanics (see Appendix A); lexical errors were not targeted.

Similar to the feedback scope, the choice of feedback strategies was influenced by the setting. Hence, direct corrections and metalinguistic codes were used; they were operationalized as defined in Ellis (2009b) and Bitchener and Storch (2016) (cf. Figure 1). That is, as in Bonilla et al. (2018), direct corrections (i.e., DCF) consisted of supplying the correct form above learners' errors. Metalinguistic codes (i.e., ME), on the other hand, placed codes (i.e., labels or abbreviations) above learners' errors to indicate their nature. As far as the codes are concerned, we employed the coding system that was used in the institution where this study took place, which added to

its ecological validity. Furthermore, it is worth noting that because the participants were already familiar with the coding system, the feedback provider (i.e., the researcher) did not need to give learners any formal training prior to the feedback intervention.

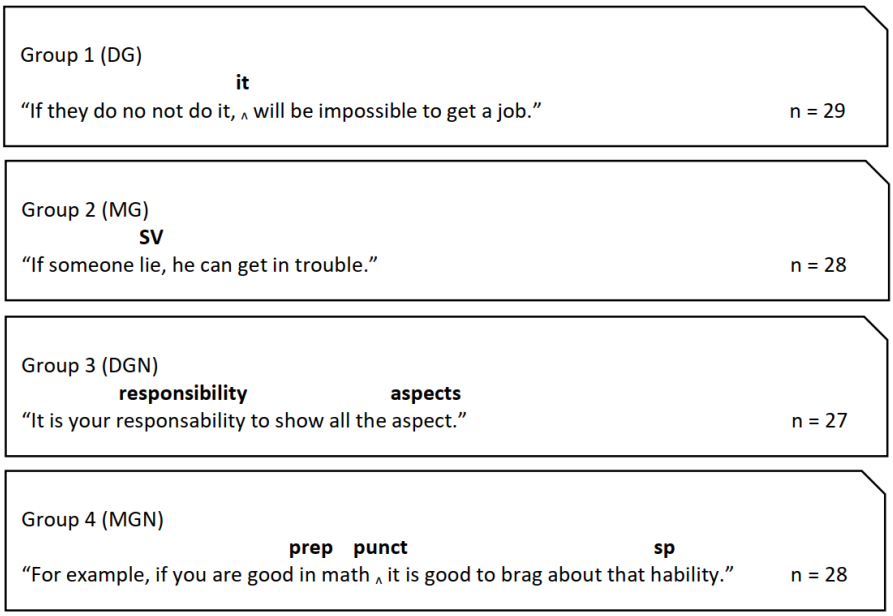

Figure 1. Implementation of error correction per experimental group. DG = direct corrections for individual grammatical issues, MG = metalinguistic codes for individual grammatical issues, DGN = direct corrections for individual grammatical and non-grammatical issues, MGD = metalinguistic codes for individual grammatical and non-grammatical issues, sv = subject-verb agreement, prep = preposition, punct = punctuation, sp = spelling

Notwithstanding this degree of familiarity, for every code that was inserted in participants' texts, both the label and its spelled-out form were written at the bottom of their essays. Contextual and theoretical reasons influenced this decision. The former relates to the fact that at the outset of the study, the participants had just returned from a 2-month vacation period. Thus, refreshing their memory was deemed desirable. The latter involves previous empirical evidence about the comprehensibility issues which codes may pose for novice EFL writers (cf. Bonilla et al., 2017). Hence, spelling out the codes was thought to counter such a potential effect. Table 1 provides a general description per error along with its abbreviated form (for metalinguistic codes).

Table 1. Error Description and Coding

Error	Brief description	Coding
Subject-verb	Subject and verb lack agreement in number	sv
Article	Unnecessary insertion, faulty, or missing definite	art
Verb	Wrong formation of verb phrase or erroneous	verb
Pronoun	Incorrect or missing pronoun	pron
Preposition	Faulty or missing preposition	prep
Word form	Faulty or missing word endings	wf
Subject deletion	Omission of subject in the sentence	sd
Subject repetition	Insertion of an unnecessary subject	sr
Sentence structure	Word order or unnecessary words or phrases	ss
Sentence fragment	Incomplete thoughts: omission of words, phrases,	frag
Spelling	Misspelled word	sp
Punctuation	Incorrect or missing punctuation mark	punct
Capitalization	Wrong or missing capitalization	cap

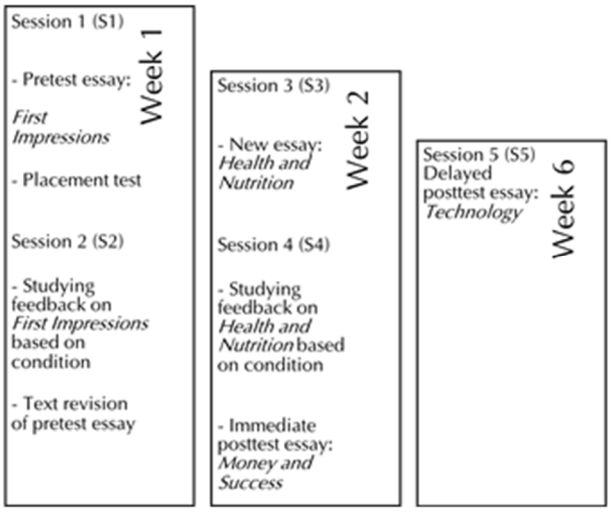

Figure 2. Overview of the procedures.

4.3 Design and procedures

A week before the semester started, the first author met with the instructors to inform them about the investigation and discuss the logistics for each work session. Then, over a 6-week period, learners participated in six writing sessions (S). Figure 2 provides an overview of the procedures. of the tasks, students had 30 minutes. After this, the first author photocopied all pretest essays.

On the first week of classes (S1), the students completed Oxford's placement test and wrote the pretest essay; for each In the case of the experimental groups, the feedback was incorporated there based on the corresponding conditions. Two days later (S2), the participants were given the photocopy of their pretest essay: those in the experimental groups were allotted 15 minutes to attend to the feedback (i.e., studying it), whereas those in the control group had the same amount of time to study their unmarked copy for self-correction purposes. For all learners in the experimental conditions, the instructions at this stage were "Study carefully the copy of the text you wrote two days ago and see in which way(s) it can be improved." Learners in the control group had the same instructions with an addition: "You can insert the changes in the copy provided."

After the time for studying the copy was over, as in Bonilla et al. (2017, 2018), learners' xeroxed essays with other-provided (i.e., the researcher) or self-provided (i.e., the learners themselves) CF were taken away because "it is obvious that a writer can look at direct corrections and copy them onto a new piece of paper" (Polio, 2012, p. 377). Also, to avoid feedback memorization issues as they have occurred in previous three-stage feedback studies (e.g., Santos, López-Serrano, & Manchón, 2010), the instructions purposely avoided hinting at the revision stage that would be next.

For the revision stage (S2), all learners were given a blank sheet and their original (unmarked) pretest essay to write a new version of their text based on what they had studied earlier. The instructions for all participants (irrespective of their condition) were the following: "Considering what you studied earlier in the copy of your composition, improve the text by writing a new version. Revise the composition using the original draft as a guide. Write it on a separate sheet". In this respect, the type of editing that students were required to do involved error-correction, deferred (on a draft), and focused on language (for a taxonomy of revision, see Stevenson et al., 2006). A week later (S3), all learners wrote a new essay, whose feedback they were able to study two days later. Immediately after that, learners did not engage in text revision of the copy they studied but wrote an immediate posttest essay (S4). Four weeks later (S5), they wrote a delayed posttest essay.

On the whole, it is worth noting that essays during all sessions were handwritten. Such a decision was contextual: while it is true that word processors could already flag many errors and, in this way, for example, significantly reduce

teacher time on specific errors and impact error treatability, a design with this variable (while empirically interesting) was beyond the contextual reality of this investigation and thus its scope. In addition, the researcher was the one who administered and supervised all sessions. In this case, the researcher's role was reduced to that of a facilitator and supervisor for timekeeping and transitions to different phases in the set-up and procedure rather than an actual instructor. To ensure that all phases were exactly the same in all groups (and hence the treatment was delivered in the same way in all groups), the first researcher prepared checklists with the different phases and their timing in order to keep track. They were checked and filled in. In addition, the researcher was provided with the class lists of all the courses involved in order to keep track of attendance.

4.4 Materials

4.4.1 Placement test

Oxford's quick placement test (OQPT) was administered to ascertain learners' proficiency level. The estimated duration of this paper-and-pen version is from 30 to 40 minutes.

4.4.2 Writing tasks

The fact that this study took place in an actual L2 writing class was decisive in deciding upon the nature of the writing task, the length of the essays, and the topics. Thus, based on the curriculum and over the course of six weeks, learners were instructed to write four 175-word argumentative essays on chapter-related topics (cf. Figure 1). The way the essays were elicited was the same in all the writing tasks; that is, the instructions consisted of a *do you-agree* question to probe learners' opinion about the topic and the same follow up question in all cases to encourage them to further elaborate on the topic (i.e., *Why or why not? Explain your reasons clearly. Use examples from your own experience to support your general ideas*). Furthermore, learners' incentive for participating in the study and engaging with the writing tasks was neither money (e.g., Storch & Wigglesworth, 2010) nor grades (e.g., Vyatkina, 2010) but awareness that because the topics were chapter-based, their essays could potentially become drafts of a graded version if their instructor deemed desirable in the future (e.g., Bonilla et al., 2017). This means that while the instructors of the course did not take part in any stage of the data collection process, they did have the option (on request) to obtain copies of the essays for further class work. As in previous work (e.g., Bonilla et al., 2018; Lavolette, Polio, & Kahng, 2015; van Beuningen et al., 2012), the tasks were not counterbalanced. In the present study, the rationale for such a decision was contextual: the topics had to be introduced in accordance with the course syllabus; hence, their order could not be altered.

4.5 Coding and analysis

To obtain a verbatim digital version of the handwritten essays, the first author and a research assistant converted them to a digital version by using a speech recognition software (i.e., Dragon Naturally Speaking 11.0). All errors were inserted manually to ensure that the handwritten essays and their digital version were identical. After that, the first author blindly coded the essays to obtain error counts of the targeted grammatical (i.e., preposition, article, subject-verb agreement, word form, verb, subject deletion, subject repetition, sentence structure, pronouns, fragment) and non-grammatical (i.e., capitalization, spelling, and punctuation) categories per session. Measures in the analyses will thus be error counts for all targeted grammatical and non-grammatical categories. More specifically, in the analyses on revision (with one analysis per error type) each individual error in the original text is a separate case. The response variable in these analyses is binary, with possible value 1 (error corrected in the revision) and 0 (error not corrected in the revision), and what we model is the probability of the error being corrected in the revision. In the analyses on independent essays (again one analysis per error type), on the other hand, cases are individual essays and the response variable is the number of errors in a text*, normalized for text size (error counts were divided by the number of words in the text, and then multiplied by 100). What we model is the evolution of these normalized error counts across consecutive new essays (i.e. going from S1 to S3 to S4 to S5).

As Murphy and Roca de Larios (2011) state, total error counts that, on the one hand, do not account for errors that were not corrected but eliminated from the text or that, on the other hand, include errors that received no feedback in an initial text, may not accurately depict error responsiveness to CF in revised essays. Therefore, in the particular case of the revision session (S2), the error counts involved first tracing each error that had received CF to determine whether it had been successfully corrected or not. For example, the preposition errors category had counts for text revision behaviors such as preposition errors corrected and preposition errors maintained. This was true for all individual targets. This approach was adopted to get a clearer picture of the amenability to correction of each category. After that, an independent experienced rater coded ten percent of the required data for interrater reliability (70 essays, randomly chosen from the five sessions [14 per session]). Table 2 shows Cronbach's alpha values for the dependent variables used for the data analyses in all conditions; all values reached acceptable reliability, that is, greater than .70 (for a review, see Taber, 2017).

Once the error counts per error category and session were completed, we proceeded to run the statistical analyses. First, we used logistic regression to examine the feedback effect during text revision (RQ1).

Table 2. Cronbach's Alphas for Interrater Reliability

Error category	
Preposition	.958
Article	.982
Subject-verb agreement	.988
Word form	.979
Verbs	.966
Subject deletion	.996
Subject repetition	.981
Sentence structure	.965
Pronouns	.985
Fragment	.958
Punctuation	.955
Spelling	.996
Capitalization	.995

Thus, with logistic regression, the contribution of the independent variable (i.e., written CF conditions) in the dependent variable (i.e., the number of correction successes versus the number of correction failures) could be measured. All logistic regression models were performed in R with the function *glm* in R package *stats* (R core team, 2016). Post-hoc comparisons (all-pair Tukey comparisons) were calculated with the function *glht* from the *multcomp* package (Hothorn et al., 2008); effect size measures using both Cox and Snell's R^2 and Nagelkerke's R^2 were calculated with the function *r2* from the *sjstats* package (Ludecke, 2017). After that, mixed-effect models were employed to examine the feedback effect on independent essays (RQ2). To this purpose, we first calculated error percentages per individual linguistic category and session; these constituted normalized versions of error counts in which low scores and high scores correspond to 'a few errors' and 'many errors', respectively. The numbers are then the normalized versions of error counts calculated with the following formula: number of errors per linguistic category/the number of words per text * 100' in line with Diab (2015). Hence, we adjusted for text length (number of words).

Mixed-effect models with random slopes and random intercepts were chosen because we have nested data, which means that the tests are nested within students. Besides, when compared with traditional ANOVAs and repeated

measures ANOVAs for the analysis of repeated measures and other types of grouped data, mixed-effect models constitute a more sophisticated alternative (Galwey, 2007; Quené & van den Bergh, 2004). All mixed-effect linear models were performed in R with the function *lmer* in R packages *lme4* (Bates, Mächler, Bolker, & Walker, 2015) and *lmerTest* (Kuznetsova, Brockhoff, & Haubo, 2016). Post-hoc comparisons (all-pair Tukey comparisons) were calculated with the function *glht* from the multcomp package (Hothorn et al., 2008); effect size measures using both R^2 and Ω^2 were calculated by the function *r2* from the *sjstats* package (Ludecke, 2017). All effect plots in this text were generated with functions from the R package *effects* (Fox, 2003).

5. Results

After presenting the preliminary analyses, this section reports the regression analyses and the mixed-effect models that were run to examine learners' successful correction of separate grammatical and non-grammatical structures during text revision (RQ1) and accurate use of separate structures on independent essays (RQ2), respectively.

5.1 Preliminary analyses

At the outset of this study, we did not find initial differences in English proficiency level (as measured by the OQPT), $F(4,139) = 1.864$, $p = .120$, $\eta_p^2 = .05$); overall grammatical accuracy, $F(4,139) = .386$, $p = .818$, $\eta_p^2 = .01$; or overall non-grammatical accuracy, $F(4,139) = .711$, $p = .586$, $\eta_p^2 = .02$ (as obtained from the pre-test essay).

5.2 Revision of essays with CF

Table 3 displays the regression analyses of correction success per response variable. The analyses did not reveal a statistically significant main effect for condition on sentence fragment, subject repetition, and verb errors in revised essays, meaning that the treatment did not have any effect on the response during text revision. We did find a statistically significant effect for condition on article, preposition, pronoun, subject deletion, sentence structure, subject-verb agreement, word form, capitalization, punctuation, and spelling errors. This implies that the short-term accuracy gains observed from an initial text to its revised version were dependent on feedback type.

Table 3. Regression Analysis of Correction Success per Error Category

	Condition					[a]CEFR_s				
	Deviance	df	Residual	Residual *df*	p	Deviance	df	Residual	Residual *df*	P
Grammatical										
Article	20.58	4	48.68	37	0.00	7.80	1	48.68	37	0.00
Sentence	3.13	4	20.04	13	0.53	0.00	1	20.04	13	0.95
Preposition	22.53	4	103.06	72	0.00	1.28	1	103.06	72	0.25
Pronoun	13.73	4	66.87	44	0.00	1.87	1	66.87	44	0.17
Subject	23.15	4	58.71	45	0.00	10.59	1	58.71	45	0.00
Subject	2.46	4	20.21	8	0.65	0.24	1	20.21	8	0.62
Sentence	37.75	4	64.43	56	0.00	5.27	1	64.43	56	0.02
Subject-verb	23.38	4	60.62	45	0.00	2.52	1	60.62	45	0.11
Verb	8.70	4	14.45	13	0.06	0	1	14.45	13	0.98
Word form	58.146	4	136.3	90	0.00	2.99	1	136.3	90	0.008
Non-grammatical										
Spelling	85.71	4	133.83	100	0.00	0.362	1	133.83	100	0.54
Punctuation	82.93	4	205.09	124	0.00	0.053	1	205.09	124	0.81
Capitalization	28.43	4	34.35	32	0.00	0.01	1	34.35	32	0.91

[a] CEFR_s = learners' standardized Common European Framework (CEFR) scores.

Table 4 summarizes the significant contrasts per error category. As can be seen, where significant differences were found in individual grammatical error categories, the DG condition significantly outperformed the self-correction group in all of them but one (i.e., article errors). Also, for errors in sentence structure, word form, subject deletion, and subject-verb agreement, the feedback treatment with direct corrections in general was significantly more advantageous than metalinguistic codes. In fact, the latter yielded evidence of superiority for one grammatical error category (i.e., subject-verb agreement) and over the self-correction group only.

Table 4. Significant Tukey Comparisons for Correction Success in Revised Essays

Error category	Condition	p	SE
Grammatical			
Article	DG > MG	0.031	1.074
Pronoun	DG > C	0.063	1.0164
Subject deletion	DG > C	0.012	1.449
	DG > MG	0.042	0.837
Sentence structure	DG > C	0.001	1.232
	DGN > C	0.027	1.156
	DG > MG	0.016	1.010
	DG > MGD	0.002	0.961
Subject-verb agreement	DG > C	0.016	1.213
	MGD > C	0.008	1.315
Word form	DG > C	< 0.001	0.594
	DGN > C	0.011	0.571
	DG > MG	< 0.001	0.489
	DG > DGN	0.02	0.457
	DG > MGD	< 0.001	0.472
	DGN > MG	0.015	0.455
Preposition	DG > C	0.010	1.111

Non-grammatical[a]			
Punctuation	DGN > C	0.002	0.324
	MGD > C	< 0.001	0.339
	DGN > DG	< 0.001	0.326
	MGD > DG	< 0.001	0.337
	DGN > DG	< 0.001	0.335
	MGD > MG	< 0.001	0.351
Spelling	DGN > C	< 0.001	0.628
	MG > C	< 0.001	0.576
	DGN > DG	< 0.001	0.514
	MGD > DG	< 0.001	0.449
	DGN > DG	< 0.001	0.502
	MGD > MG	0.004	0.435

Note. DG = direct corrections for individual grammatical issues, MG = metalinguistic codes for individual grammatical issues, DGN = direct corrections for individual grammatical and non-grammatical issues, MGD = metalinguistic codes for individual grammatical and non-grammatical issues, C = self-correction (without feedback provision).

[a]Within this category, the model for capitalization errors yielded a significant condition effect (cf. Table 2). However, none of the post-hoc comparisons reached statistical significance, which could be the result of multiple comparisons taking away some of the power. Therefore, capitalization errors are not included in the table.

From Table 4 it can also be observed that separate non-grammatical issues proved responsive to written CF and that both direct corrections and metalinguistic codes could effectively treat them as no statistically significant differences were observed between the two feedback types. In addition, it also reveals that the groups which received written CF on separate non-grammatical issues significantly outperformed those groups which did not (i.e., the self-correction condition).

5.3 Independent essays

The mixed-effect model did not reveal a statistically significant interaction effect of time and condition for fragment ($X^2_4 = 3.818$, $p = 043$, $R^2 = .66$, $\Omega^2 = .57$), pronoun ($X^2_4 = 5.878$, $p = 0.208$, $R^2 = .47$, $\Omega^2 = .43$), subject deletion ($X^2_4 = 2.949$, $p = 0.566$, $R^2 = .57$, $\Omega^2 = .47$), subject repetition ($X^2_4 = 0.701$, $p <= 0.951$, $R^2 = .42$, $\Omega^2 = .36$), subject-verb agreement ($X^2_4 = 5.914$, $p = 0.205$, $R^2 = .31$, $\Omega^2 = .25$), verb ($X^2_4 = 3.538$, $p = 0.472$, $R^2 = .32$, $\Omega^2 = .26$), and spelling ($X^2_4 = 5.712$, $p = 0.221$, $R^2 = .58$, $\Omega^2 = .54$) errors. These results indicate that the treatment did not play a role in the reduction of these error types on independent essays.

A statistically significant interaction effect of time and condition was found for article ($X^2_4 = 14.02$, $p = 0.007$, $R^2 = .56$, $\Omega^2 = .47$), preposition ($X^2_4 = 13.72$, $p = 0.008$, $R^2 = .31$, $\Omega^2 = .26$), sentence structure ($X^2_4 = 14.23$, $p = 0.006$, $R^2 = .38$, $\Omega^2 = .34$), word form ($X^2_4 = 9.608$, $p = 0.047$, $R^2 = .48$, $\Omega^2 = .45$), capitalization ($X^2_4 = 21.69$, $p = 0.000$, $R^2 = .45$, $\Omega^2 = .34$), and punctuation ($X^2_4 = 21.88$, $p = 0.000$, $R^2 = .64$, $\Omega^2 = .59$) errors. This means that the decrease over time in the error score of these categories hinged on feedback type. Table 5 displays the descriptive statistics of error frequency on independent essays. Table 6 provides an overview of significant Tukey comparisons per error category. As can be gleaned, only direct corrections had an effect beyond text revision.

Table 5. Descriptive Statistics: Number of Errors per 100 Words

	DG		MG		DGN		MGD		C	
	M	SD	M	SD	M	SD	M	SD	M	SD
Grammatical										
Article										
Session 1	.19	.412	.325	.416	.325	.611	.263	.724	.268	.490
Session 3	.47	.591	.169	.335	.311	.413	.230	.383	.252	.463
Session 4	.24	.475	.236	.435	.233	.537	.251	.455	.661	.901
Session 5	.15	.346	.238	.341	.116	.289	.405	.903	.625	.626
Preposition										
Session 1	.74	.598	.393	.355	.553	.774	.559	.658	.317	.594
Session 3	.39	.501	.468	.533	.538	1.19	.598	.605	.254	.494
Session 4	.28	.437	.239	.379	.122	.286	.554	.603	.454	.631
Session 5	.26	.425	.209	.381	.227	.447	.371	.490	.482	.522
Sentence structure										
Session 1	.33	.430	.255	.384	.426	.519	.460	.800	.478	.497
Session 3	.63	.700	.534	.746	.370	.584	.216	.714	.526	.507
Session 4	.08	.216	.220	.336	.466	.618	.402	.570	.480	.581
Session 5	.06	.208	.378	.465	.301	.499	.616	.759	.796	.627
Word form										
Session 1	1.0	1.10	1.07	1.13	.920	.923	.962	1.16	.826	1.07
Session 3	1.0	109	.869	.968	1.51	1.45	1.17	1.24	.964	1.04
Session 4	.60	.691	.641	1.09	.631	.721	1.17	1.74	.965	1.05
Session 5	.34	.412	.712	.708	.544	.753	.620	.837	1.04	.777

Non-grammatical										
Capitalization										
Session 1	.11	.293	.077	.172	.560	.652	.199	.414	.198	.482
Session 3	.32	.611	.188	.345	.351	.505	.492	.520	.351	.495
Session 4	.28	.493	.327	.495	.066	.199	.310	.434	.380	.447
Session 5	.28	.452	.225	.477	.119	.300	.155	.335	.313	.462
Punctuation										
Session 1	2.0	1.34	1.87	1.43	2.66	1.66	2.02	1.12	1.85	1.36
Session 3	2.0	1.32	1.43	1.32	2.46	1.28	2.62	1.82	1.97	1.54
Session 4	2.5	1.79	1.86	1.40	1.84	1.28	2.24	1.57	2.61	1.71
Session 5	2.7	1.62	1.98	1.27	1.46	1.92	2.35	1.51	3.00	1.46

Note. DG = direct corrections for individual grammatical issues, MG = metalinguistic codes for individual grammatical issues, DGN = direct corrections for individual grammatical and non-grammatical issues, MGD = metalinguistic codes for individual grammatical and non-grammatical issues, C = self-correction (without feedback provision).
The error counts are calculated with Diab's (2015) formula: number of errors per linguistic category/the number of words per text * 100.

Table 6. Significant Tukey Comparisons for Error Reduction on Independent Essays

Error category	Condition	p	SE
Grammatical			
Article	DG > C	0.031	0.063
	DGN > C	0.007	0.064
Prepositions	DG > C	0.011	0.066
	DGN > C	0.013	0.067
Sentence structure	DG > C	0.003	0.062
Word form	DG > C	0.004	0.110
Non-grammatical			
Punctuation	DGN > C	< 0.001	0.182
	DGN > DG	0.002	0.181
Capitalization	DGN > C	0.002	0.055
	DGN > DG	0.001	0.055
	DGN > MG	< 0.001	0.055

Note. DG = direct corrections for individual grammatical issues, MG = metalinguistic codes for individual grammatical issues, DGN = direct corrections for individual grammatical and non-grammatical issues, MGD = metalinguistic codes for individual grammatical and non-grammatical issues, C = self-correction (without feedback provision).

6. Discussion

Running counter to previous claims (e.g., Sheen, 2007), we did not find any evidence that comprehensive CF taxed learners' ability to process the feedback to the extent that they could not show short- and long-term accuracy gains of separate grammatical and non-grammatical constructions. From a cognitive perspective (see Bitchener & Ferris, 2012; Bitchener, 2016 for a thorough discussion), this means that learners had enough attentional resources to attend to the feedback, engage in a cognitive comparison, modify their output, and retrieve on independent essays any explicit knowledge they had gained from studying the feedback and putting it in practice during essay revision. However, the results also indicate that (1) the targeted structures benefited differently from the feedback types (i.e., feedback type and error type proved to be influential variables), (2) the effect of direct corrections and metalinguistic codes was not equally durable (i.e., direct corrections were superior), (3) a less comprehensive treatment proved more effective for treating individual grammatical issues (i.e., feedback scope played a role), and (4) learners did not have accuracy gains in error categories they did not receive feedback on (i.e., there was no evidence of learning where noticing was not required).

Specifically, the research questions dealt with the differential effects of direct corrections and metalinguistic codes to bring about not only successful error correction of separate grammatical and non-grammatical errors when revising essays after receiving CF (RQ1) but also subsequent error reduction on independent essays (RQ2). As in other studies (e.g., Bitchener et al., 2005; Ferris & Roberts, 2001), our evidence indicates that there were differences across error types. In this respect, Figure 3, which summarizes our results into patterns of response, allows a number of theoretically based observations. First, from the group of structures that yielded both significant correction success in revision and error frequency decrease over time, three error types (i.e., sentence structure, word form, and preposition) have been referred to as untreatable or too complex to treat (e.g., Ferris, 1999). For instance, out of three targeted categories (i.e., articles, past tense, and prepositions), Bitchener et al. (2005) found evidence of L2 development in all of them but one, that is, in prepositions. However, one aspect in common between Bitchener et al. (2005) and our study is that both confirm that one particular error category (i.e., articles) is a highly correctable target under different conditions (see also Bitchener & Knoch, 2010b; Nassaji, 2011; Shintani et al., 2014). Second, from the group of structures that yielded significant correction success in revision but no error frequency decrease over time, an error category such as pronouns had significant accuracy gains in an immediate posttest in Diab (2015), but in the present study, the effect did not last beyond revision. There is a similarity, however, between our study and Ferris (2006), where subject-verb agreement errors proved responsive to written CF in a revised text but showed no sustained effect over time.

Third, from the group of structures with no feedback responsiveness at all, two error categories (i.e., verb and fragment) rendered significant progress from an initial text to its revised version in Ferris (2006), which was not evident in our results. Hence, considering the aforementioned variation in amenability to correction, the potential sources of influence on error treatability may deal with issues beyond the rule-based nature of an error—running counter to Ferris's treatable/untreatable dichotomy.

What is more, the feedback effectiveness differed across treatments as well. Grammatical errors, for example, responded differently: syntactic (e.g., subject deletion and sentence structure) and morphological (e.g., word form) issues were exclusively amenable to treatment with direct corrections. In addition, where significant differences were found in long-term error reduction, the findings suggest that direct corrections may have the upper hand. For instance, any short-term effect that metalinguistic codes had on non-grammatical errors wore off beyond text revision, and after feedback provision, only direct corrections still had a durable effect for grammatical errors (e.g., preposition, article, sentence structure) and non-grammatical ones (e.g., capitalization and punctuation). Nevertheless, where significant differences were found in correction success during text revision, both direct corrections and metalinguistic codes were equally successful in enabling learners to self-edit all error categories in the non-grammatical type (e.g., spelling and punctuation).

Consequently, the analyses partially support our expectations (cf. 3.2): separate grammatical errors indeed become better targets with direct corrections, but non-grammatical ones do not respond better with metalinguistic codes in the long run.

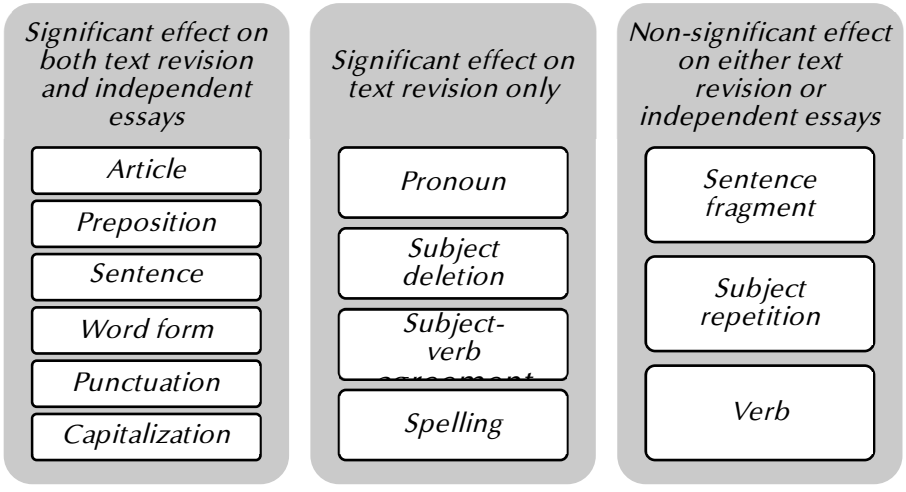

Figure 3. Patterns of response of individual error categories.

This means that the expected clear-cut amenability to correction, which had been originally predicted based on van Beuningen et al.'s (2012) overall accuracy findings, may not entirely apply in a study on individual structures. Instead, our findings reveal an intricate interplay of variables which corroborate the "complexity of corrective feedback and how its effectiveness may interact with various factors" (Lira-Gonzales & Nassaji, 2020, p. 18). Indeed, the present empirical work renders seven variables which could be considered as influential factors in the amenability to correction of individual features (cf. Figure 4). Out of them, four constitute an addition to the literature, whereas three—previously identified—further consolidate themselves as variables that could play a role in error treatability under other conditions as well (e.g., feedback provided comprehensively to FL writers).

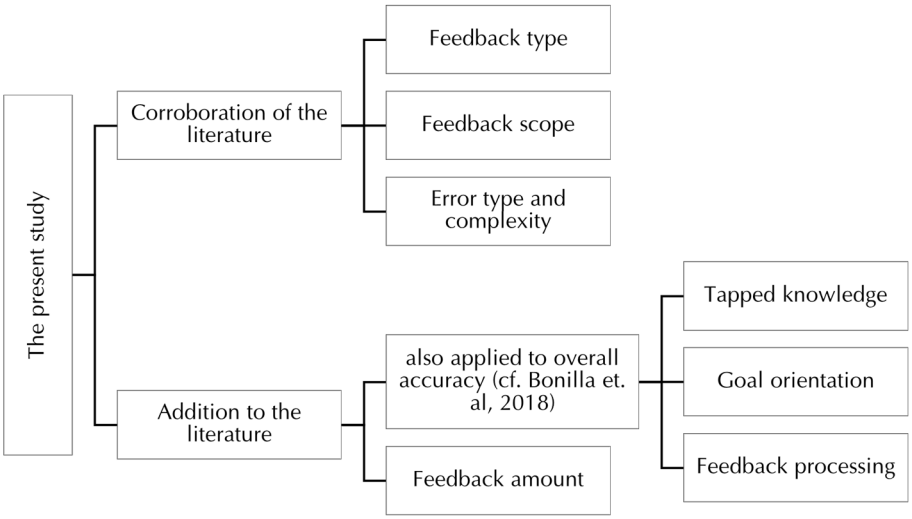

Figure 4. Influential factors in error treatability

Against the aforementioned, what follows is a discussion of the attributing factors in Figure 4. They will be presented in a way that differences in accuracy performance and pedagogical/theoretical implications are touched upon.

TAPPED KNOWLEDGE = Separate grammatical and non-grammatical accuracy gains were in line with the knowledge that was tapped in the conditions.

One advantage of the current analyses is that they provide a detailed look at error treatability in ways that overall accuracy measures cannot. As a result, a valuable contribution of the present study is the evidence that reveals that variables that may explain overall accuracy gains could also be attributed to the correctability of individual features (see also Goal orientation and Feedback processing). For

example, in agreement with the transfer appropriate principle (Lightbown, 2008) and as in Bonilla et al. (2018), no condition yielded evidence of grammatical or non-grammatical improvement in individual targets that learners were not required to attend to. In other words, accuracy improvement of individual targets only occurred in groups whose attention was drawn to their grammatical issues (i.e., DG, MG, DGN, MGD), and non-grammatical accuracy improvement took place only in groups whose attention was drawn to their non-grammatical inaccuracies (i.e., DGN, MGD). To this day, a "match between the processes and conditions that are present during learning and those that are present at the time of retrieval" (Lightbown, 2008, p. 42) had not been pinpointed as a variable that could determine error treatability, hence, its theoretical and pedagogical significance. On the one hand, evidence on the effectiveness of written CF may not paint an accurate picture if, for example, the learning conditions involve feedback on a composition and the retrieval conditions involve an error detection/correction test (e.g., Asassfeh, 2013). On the other hand, L2 (teachers) composition teachers might want to reflect both on their feedback practices and on the tasks they administer in order to determine whether or not they match the L2 learning goals they have set for their students.

GOAL ORIENTATION = The grammatical and non-grammatical accuracy gains also matched the knowledge that was triggered by the revision instructions.

The revision instructions, which instructed learners to revise the text based on other- or self-provided feedback, may have triggered the knowledge base that the conditions had already tapped; this may have in turn influenced learners' goal orientation for text revision. Similar evidence has been found in Wallace and Hayes (1991), who investigated the relation between revision and task definition. The authors found that a specific 8-minute instruction to revise globally helped college learners to improve a draft more than a general cue to make the text better—a finding later confirmed in Wallace et al. (1996). Thus, it is plausible that with revision instructions such as *considering what you studied earlier in the copy of your composition, improve the text by writing a new version*, learners in our study may have fine-tuned their accuracy goals towards revision changes involving the knowledge that the experimental conditions had provided (i.e., grammatical only or both grammatical and non-grammatical). This could help explain why there were no accuracy gains in knowledge that was neither tapped (in the conditions) nor triggered (by the instructions). Such results could then be a confirmation of the fact that when teaching higher-education students to revise, a goal may suffice (cf. Van Steendam, Rijlaarsdam, & Van den Bergh, 2018). Nevertheless, given the novelty of this factor in error treatability literature, more research is warranted to validate this interpretation.

FEEDBACK TYPE AND PROCESSING = Error categories thought to be complex were responsive to written CF.

Drawing on Ferris's (1999) treatable/untreatable distinction, errors in sentence structure, prepositions, and word form have typically been considered untreatable (i.e., complex and idiosyncratic). However, in the present study, not only were errors in those categories successfully corrected in a revised draft but also their frequency significantly reduced on independent essays. Such responsiveness to feedback treatment does not corroborate the results in Bitchener et al. (2005) and Nassaji (2011) for prepositions or in Ferris (2006) for sentence structure errors. A likely cause for this discrepancy could originate from the feedback strategies that were employed (i.e., feedback type) and what learners were required to do with the feedback (i.e., feedback processing). Frear and Chiu (2015) explain, for example, that a written CF strategy could be either input providing when learners receive the correct forms (e.g., with direct corrections) or output prompting (also output pushing) when learners correct on their own without explicit provision of the correct forms (e.g., with metalinguistic codes). Other researchers acknowledge this dichotomy (e.g., Ellis, 2010; Sheen, 2010; Shintani et al., 2014) and note that such labeling is not fixed, meaning that it also hinges on what exactly learners do with the feedback. Therefore, in this study, an input-providing strategy such as direct corrections could have also become output prompting because the feedback processing entailed both studying the feedback and revising a text without access to it. Thus, given that the nature of direct corrections itself facilitates an immediate cognitive comparison (Ellis, 2010) and that the feedback processing in the present study involves pushed output, its effect may have been magnified so much so that it was able to bring about significant accuracy changes (and outperform metalinguistic codes) even in error categories otherwise construed as 'untreatable.' Therefore, a unique feature of this study is twofold: the confirmation that feedback type has a role to play in the extent to which a separate error category is correctable and the suggestion of a maximized feedback effect as a result of what FL learners are asked to with the input received (e.g., in the form of written CF).

FEEDBACK AMOUNT = Not all the error categories received an equal amount of feedback.

How much feedback learners receive may be directly linked to how many errors they make on an initial writing piece to begin with (cf. Truscott, 2001). The saliency of the feedback in the input (e.g., written CF) has only been recently hinted at as a potential factor that may determine the extent to which an error correction is noticeable, influencing in turn the extent to which learners attend to it (Shintani et al., 2014)—hence the relevance of this addition to the literature. For example, those authors found that learners significantly improved the accuracy of the structure which they had received numerous corrections on (i.e., the hypothetical conditional) and failed to do so with the least salient structure in the feedback (i.e.,

the indefinite article). Similarly, in our study, the amount of feedback on a given individual category was not equal as learners had more errors on some categories than on others. For example, the three error categories that had no feedback response at all (cf. Figure 3) were precisely the ones that received the least amount of feedback due to the low frequencies on learners' independent essays (cf. Appendix B). Hence, it should not be surprising that in those cases, learners were not able to retrieve on independent essays the knowledge that they learned in the feedback sessions: the explicit knowledge that was learned was insufficient to be subsequently applied in new contexts. This could imply that consolidation of L2 knowledge may not occur unless learners are given sufficient input and opportunities for repeated retrievals as Bitchener and Storch (2016) suggest.

FEEDBACK SCOPE = Non-grammatical issues were effectively treated even when attention was also drawn to grammatical ones, but grammatical error categories responded significantly better with a less comprehensive CF treatment that did not include attention to non-grammatical issues.

Feedback scope within a comprehensive approach to errors has largely been an underresearched variable. The only two studies with a comprehensive CF group in their baseline comparison have been Frear (2010) and Frear and Chiu (2015), who compared comprehensive CF with highly selective CF. Although their results indicate that attention to a large number of errors does not hinder learners' ability to improve the accuracy of one targeted feature (i.e., past tense verbs), their design is not directly comparable to ours. Thus, the comparison of different comprehensive CF forms in the present study widens current knowledge by addressing underexplored attentional issues such as how feedback learners are able to handle (as evidenced in their individual grammatical and non-grammatical accuracy improvement). The evidence from the extended analyses not only further consolidates the role of feedback scope in enhancing (or not) error correctability but also provides theoretically and pedagogically relevant evidence. For example, similar to Bonilla et al. (2018) our findings suggest that the accuracy of separate non-grammatical targets (e.g., punctuation) may not be affected if attention is also paid to grammatical ones. However, the accuracy with which learners improve their use of some grammatical structures (e.g., word form) may only be further maximized when learners' attention is drawn solely to grammatical issues.

ERROR TYPE = Individual grammatical and non-grammatical issues responded differently to direct corrections or metalinguistic codes.

Drawing on van Beuningen et al.'s (2012) conclusion that overall grammatical and non-grammatical accuracy may be more responsive to direct corrections and metalinguistic codes, respectively, the hypotheses in our study expected the same response in the targeted individual error categories. Nonetheless, the results partly

corroborate those of overall accuracy in van Beuningen et al. (2012): individual non-grammatical targets (i.e., spelling, punctuation, and capitalization) did not do significantly better with metalinguistic codes but with direct corrections. This contradiction brings into question not only the comparability of the studies given their differences (among others) in accuracy measures but also previous discussions of error type in light of the treatable/untreatable dichotomy—shedding in turn some light into the theoretical discussion of error treatability. Indeed, it suggests that error amenability to written CF may not hinge entirely on error type but on the complexity of a given error irrespective of its type.

ERROR COMPLEXITY = Errors within the same type (e.g., grammatical or non-grammatical) responded differently to written CF.

Because of the lack of theoretical basis in the treatable/untreatable distinction (Shintani et al., 2014; van Beuningen, 2010), error complexity has been pointed out as a more reasonable factor to explain differences in error treatability (e.g., Shintani et al., 2014). The results in the present study lend support to this stand and raise the need to discuss what particular characteristics could make an error more complex (or not), especially in instances when feedback amount seems not to be decisive in bringing about L2 development. The possibility that complexity may differ in error categories within the same type (e.g., grammatical or non-grammatical) was seen, for example, in targets that received slightly more feedback (e.g., pronoun, subject deletion, subject-verb agreement, and spelling) than others (e.g., articles, punctuation, and capitalization) and yet failed to sustain the feedback effect beyond revision. While not claiming to be a comprehensive overview, Table 7 seeks to elucidate potential complexity issues by taking a closer look at the knowledge that Spanish L1 EFL learners would need to grasp to apply the L2 rules correctly in new contexts.

As can be seen in Table 7, more choices are involved in expressing the intended meaning with the correct form in the correct place for *subject-verb agreement* and *pronoun* (a problem of form in DeKeyser, 2005) than for *articles*. In addition, while subject-deletion issues may arise in EFL learning due to the null subject in Spanish L1 (a problem of meaning due to novelty in DeKeyser, 2005), that may not be the case with article usage, whose system in Spanish and English is similar. Therefore, if "[o]ne way correctability can be judged is by the extent of the practical problems involved in correcting each error type" (Truscott, 2001, p. 94), the seemingly more transparent rules for articles may have enabled learners to apply them and retrieve them significantly better than the potentially opaque rules for pronoun and subject deletion. That is, Kiparsky (1971) distinguishes grammatical rules by difficulty: transparent rules are those that are easy (to grasp) whereas opaque rules are those that are hard.

Table 7. Overview of Potential Sources of Error Complexity

Error	Knowledge required for proper usage in English L2
Grammatical	
Pronoun	Morphological distinctions (case, person, gender, and
	Contextual information (referent)
Subject deletion	Parts of the sentence
	The optionality of the subject (i.e., the null subject is
Subject-verb agreement	The principle of grammatical concord
	The principle of notional concord
	The principle of proximity
Articles	Noun classes
	Indefinite and definite reference
Non-grammatical	
Spelling	British- or American-oriented subsystems
	Morphological, phonological, orthographic form
	Assimilation of foreign words
Punctuation	Clause/sentence boundaries
	British- or American-oriented subsystems
Capitalization	Referent identification
	Noun classes

Note. Summary based on the reference manual *A Comprehensive Grammar of the English Language* (Quirk, Greenbaum, Leech, & Svartvik, 1985).

Further, the same criteria used to judge potential error complexity of grammatical errors (i.e., practical problems as in Truscott, 2001 or error characteristics as in DeKeyser, 2005) could be applied for non-grammatical ones. For example, Truscott (2001) states that as part of the practical problems in error treatability, *discreteness* could make a difference. The author defines this criterion as the need to "deal with a given item in a variety of contexts" (p. 99). Along these lines then, error treatability of spelling may be more complex than punctuation and capitalization given that feedback on spelling is clearly "bound to the error's original text" (Truscott, 2001, p. 95), rendering it harder to apply in a new context as evidenced in the lack of significant improvement over time. However, given the scant evidence on error

treatability of individual non-grammatical targets, whether or not being thematically bound makes spelling more opaque than punctuation and capitalization is in need of further scrutiny. In this sense, the present study offers a springboard for more research to take place.

7. Limitations and future work

The following limitations and suggestions should be kept in mind for a future research agenda. To begin with, adding to the ecological validity of this study, the writing tasks were curriculum-based and not restricted to elicit a (pre-selected) narrow number of error categories. While this could be considered a strength in the study because it gives learners the chance to express their ideas freely (see Bruton, 2009), it could also be a limitation because topic differences (albeit the elicitation was the same in all tasks) may not have afforded enough context to generate the same structures across testing times. Most likely, a more controlled series of tasks would have resolved this issue. On this point, one step further to take advantage of authentic L2 writing tasks (and which we did not attempt) is incorporating measures of linguistic complexity and lexical diversity, which remain underresearched to this day. The only study on comprehensive CF that has addressed three measures (grammatical accuracy, structural complexity, and lexical diversity) is van Beuningen et al. (2012) with their sample consisting of SL Dutch learners. However, whether the same results could be obtained with EFL writers is unknown. In like manner, the present study was conducted with English (Teaching) majors in a FL environment, so the generalizability of its findings with other learner types in other settings warrants further investigation. Along the same lines, because the analyses considered a great number of outcomes (multiple error types by 4 conditions) with a relatively small sample, caution should be exercised in drawing conclusions.

Also, the instructional context of this study did not allow us to sustain a no-feedback condition for more than eight weeks. Thus, EFL classroom-based feedback studies that examine a more longitudinal period remain in order. A related need are studies that provide feedback on more than two occasions. It would be interesting to see if the error categories that showed short-term accuracy gains but failed to do so over time (e.g., subject deletion, pronoun, spelling) could have benefited from more feedback sessions or a different treatment. Equally necessary is to determine whether the feedback effect on those that yielded a significant decrease (e.g., word form, prepositions) would have diluted in a longer term. In doing so, the relevance of examining errors in fine-grained categories remains. One possible step in this direction could be more studies exploring the treatability of errors within open (e.g., regular past tense verbs) and close (i.e., irregular past tense verbs) domains. Still related to context, while research practices that conform with teaching practices bolster the ecological validity of an investigation (e.g., feedback on

handwritten essays as is the case in this study), future research that looks into the feedback effect on error treatability when word processors are used may be worth pursuing.

It would also be interesting to explore to what extent noticing is further influenced by revision instructions. Bearing in mind that attention is primordial for L2 learning (Schmidt, 2001), if part of what learners need to further maximize the feedback effect also depends on (clearer) instructions, its implementation in L2 (writing) classes seems straightforward and the pedagogical implications would be of great value.

8. Concluding remarks and implications

The purpose of this study was to investigate the effect that CF on multiple errors has on learners' accurate use of separate grammatical and non-grammatical structures both during text revision and on independent essays. With this in mind, the findings provide answers to key theoretical and practical issues in L2 language teaching and writing. First, from our study there is now evidence that simultaneous corrections of diverse errors can lead to interlanguage development of separate features. The present study confirms that error categories that have proved amenable when treated selectively and on one occasion (e.g., articles, pronouns) can also respond well with more sustained CF (e.g., on two occasions) and in conjunction with multiple errors (e.g., non-grammatical issues). Consequently, L2 teachers and researchers can be confident that drawing learners' attention to a broad range of separate features will not prevent them from achieving immediate (grammatical or non-grammatical) accuracy or improving their use of language forms over time.

Additionally, in error detection studies, empirical evidence has already pointed at the fact that detection "varies according to the nature of the errors in the texts and the processing demands required to detect them" (Larigauderie, Guignouard, & Olive, 2020). However, when it comes to error correction, the picture has not been that clear. Now, with the present study, more support is provided to previous claims about error treatability being the result of a complex interaction of different factors (cf. Nassaji, 2011; Sheen, 2010; Shintani et al., 2014). For example, our results agree with those of Shintani et al. (2014) in that the degree of complexity of the rule involved may play a more significant role in error correctability than the rule-based nature of an error. The present study then further adds theoretical and practical knowledge to the ongoing discussion on error treatability by suggesting novel potentially influential variables such as tapped knowledge and feedback processing. Indeed, if the extent to which FL learners internalize corrections may hinge on an intricate set of factors, finding out which error types are amenable to which sort of teacher feedback has important implications for the feedback that teachers must provide and for L2 writing/revision instruction.

Finally, our findings confirm that the saliency of a structure in the input (in the form of written CF) can influence the degree to which learners pay attention to it (cf. Shintani et al., 2014). In fact, similar to the findings in Bonilla et al. (2018), learners are not likely to correct the L2 written errors they are not required to attend to. Thus, if goal orientation (as triggered by the written CF treatment in conjunction with feedback processing instructions) also plays a role in error treatability in L2 writing, accuracy improvement of individual error categories may not necessarily hinge on cognitive load issues due to scope (as suggested in Sheen et al., 2009) but on triggering a specific knowledge base. If so, the pedagogical and theoretical implications for CF writing feedback instruction are noteworthy given the pivotal role that attention is thought to play in L2 learning (cf. Schmidt, 2001), namely feedback uptake and retention.

Note

* The term "normalized frequencies"—as used in, for example, corpus linguistics and other analyses of textual data—means something completely different from "normalization of scores" as sometimes used in educational assessment (for an example in corpus linguistics, see Chapter 2 on Vocabulary by Brezina (2018)).

References

Bonilla, López M., Van Steendam, E., & Buyse, K. (2017). Comprehensive corrective feedback on low and high proficiency writers: Examining attitudes and preferences. *ITL-International Journal of Applied Linguistics*, *168*(1), 91–128. https://doi.org/10.1075/itl.168.1.04bon

Bonilla, López M., Van Steendam, E., Speelman, D., & Buyse, K. (2018). The Differential Effects of Comprehensive Feedback Forms in the Second Language Writing Class: Comprehensive Feedback in the L2 Writing Class. *Language Learning*, *68*(3), 813–850. https://doi.org/10.1111/lang.12295

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). *lme4: Linear Mixed-Effects Models Using Eigen and S4*. (Version R package version 1.1-10) [Computer software]. URL http://CRAN.R-project.org/package= lme4.

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

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

Bitchener, J. (2012b). Written Corrective Feedback for L2 Development: Current Knowledge and Future Research. *TESOL Quarterly*, *46*(4), 855–860. https://doi.org/10.1002/tesq.62

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

Bitchener, J. (2016). To what extent has the published written CF research aided our understanding of its potential for L2 development? *ITL - International Journal of Applied Linguistics*, *167*(2), 111–131. https://doi.org/10.1075/itl.167.2.01bit

Bitchener, J., & Storch, N. (2016). *Written Corrective Feedback for L2 Development* (Vol. 96). Multilingual Matters. https://doi.org/10.1075/itl.167.2.01bit

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

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

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

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

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

Brown, D. (2012). The Written Corrective Feedback Debate: Next Steps for Classroom Teachers and Practitioners. *TESOL Quarterly*, *46*(4), 861–867. https://doi.org/10.1002/tesq.63

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

Diab, N. (2015). Effectiveness of written corrective feedback: Does type of error and type of correction matter? *Assessing Writing*, *24*, 16–34. https://doi.org/10.1016/j.asw.2015.02.001

Ellis, R. (2010). A Framework for Investigating Oral and Written Corrective Feedback. *Studies in Second Language Acquisition*, *32*(02), 335–349. https://doi.org/10.1017/S0272263109990544

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

Farrokhi, F., & Sattarpour, S. (2012). The Effects of Direct Written Corrective Feedback on Improvement of Grammatical Accuracy of High- proficient L2 Learners. *World Journal of Education*, *2*(2). https://doi.org/10.5430/wje.v2n2p49

Ferris, D., & Eckstein, G. (2020). Language matters: Examining the language-related needs and wants of writers in a first-year university writing course. *Journal of Writing Research*, *12*(2), 321–364. https://doi.org/10.17239/jowr-2020.12.02.02

Ferris, D. (1999). The case for grammar correction in L2 writing classes: A response to Truscott (1996). *Journal of Second Language Writing*, *8*(1), 1–11. https://doi.org/10.1016/s1060-3743(99)80110-6

Ferris, D. (2004). The "Grammar Correction" Debate in L2 Writing: Where are we, and where do we go from here? (and what do we do in the meantime …?). *Journal of Second Language Writing*, *13*(1), 49–62. https://doi.org/10.1016/j.jslw.2004.04.005

Ferris, D. (2006). Does error feedback help student writers? New evidence on the short- and long-term effects of written error correction. In K. Hyland & F. Hyland (Eds.), *Feedback in second language writing: Contexts and issues* (pp. 81–104). Cambridge University Press. https://doi.org/10.1017/cbo9781139524742.007

Ferris, D. (2010). Second language writing research and written corrective feedback in SLA. *Studies in Second Language Acquisition*, *32*(02), 181–201. https://doi.org/10.1017/S0272263109990490

Ferris, D., & Roberts, B. (2001). Error feedback in L2 writing classes: How explicit does it need to be? *Journal of Second Language Writing*, *10*(3), 161–184. https://doi.org/10.1016/s1060-3743(01)00039-x

Field, A. (2013). *Discovering statistics using IBM SPSS statistics*. Sage.

Fitzgerald, J. (1987). Research on Revision in Writing. *Review of Educational Research*, *57*(4), 481–506.

Fitzgerald, J., & Markham, L. (1987). Teaching Children about Revision in Writing. *Cognition and Instruction*, *4*(1), 3–24. https://doi.org/10.1207/s1532690xci0401_1

Fox, J. (2003). Effect displays in R for generalised linear models. *Journal of Statistical Software*, *8*(15), 1–27.

Frear, D. (2010). The Effect of Focused and Unfocused Direct Written Corrective Feedback on a New Piece of Writing. *College English: Issues and Trends*, *3*, 59–71.

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

Galwey, N. (2007). *Introduction to Mixed Modelling: Beyond Regression and Analysis of Variance*. John Wilwy & Sons.

Guénette, D. (2007). Is feedback pedagogically correct? *Journal of Second Language Writing*, *16*(1), 40–53. https://doi.org/10.1016/j.jslw.2007.01.001

Guénette, D. (2012). The pedagogy of error correction: Surviving the written corrective feedback challenge. *TESL Canada Journal*, *30*(1), 117. https://doi.org/10.18806/tesl.v30i1.1129

Hothorn, T., Bretz, F., & Westfall, P. (2008). Simultaneous Inference in General Parametric Models. *Biometrical Journal*, *50*(3), 346–363. https://doi.org/10.1002/bimj.200810425

Hyland, F. (2003). Focusing on form: Student engagement with teacher feedback. *System*, *31*(2), 217–230. https://doi.org/10.1016/S0346-251X(03)00021-6

Kurzer, K. (2017). Dynamic Written Corrective Feedback in Developmental Multilingual Writing Classes. *TESOL Quarterly*. https://doi.org/10.1002/tesq.366

Kuznetsova, A., Brockhoff, P. B., & Haubo, R. (2016). *lmer Test: Tests in Linear Mixed Effects* (Version R package 2.0-32) [Computer software]. https://www.r-project.org/

Lira-Gonzales, M.-L., & Nassaji, H. (2020). The Amount and Usefulness of Written Corrective Feedback Across Different Educational Contexts and Levels. *TESL Canada Journal*, *37*(2), 1–22. https://doi.org/10.18806/tesl.v37i2.1333

Lavolette, E., Polio, C., & Kahng, J. (2015). The accuracy of computer-assisted feedback and students' responses to it. *Language, Learning & Technology*, *19*(2), 50–68.

Lee, I. (2013). Research into practice: Written corrective feedback. *Language Teaching*, *46*(01), 108–119. https://doi.org/10.1017/S0261444812000390

Liu, Q., & Brown, D. (2015). Methodological synthesis of research on the effectiveness of corrective feedback in L2 writing. *Journal of Second Language Writing*, *30*, 66–81. https://doi.org/10.1016/j.jslw.2015.08.011

Long, M. (1996). The role of the linguistic environment in second language acquisition. In W. C. Ritchie & T. K. Bhatia (Eds.), *Handbook of second language acquisition* (pp. 413–468). Academic Press. https://doi.org/10.1016/b978-012589042-7/50015-3

Ludecke, D. (2017). *sjstats: Statistical Functions for Regression Models* (Version R package 0.10.2) [Computer software]. https://cran.r-project.org/web/packages/sjstats/index.html

Manchón, R. (2011). Situating the learning-to-write and writing-to-learn dimensions of L2 writing. In *Learning-to-write and writing-to-learn in an additional language* (Vol. 31, pp. 3–14). John Benjamins Publishing. https://doi.org/10.1075/lllt.31.03man

Murphy, L., & Roca de Larios, J. (2011). *Feedback in Second Language Writing: An introduction*. http://digitum.um.es/xmlui/handle/10201/23409

Nassaji, H. (2011). Correcting students' written grammatical errors: The effects of negotiated versus nonnegotiated feedback. *Studies in Second Language Learning and Teaching*, *3*, 315–334.

Pawlak, M. (2014). *Error Correction in the Foreign Language Classroom*. Springer Berlin Heidelberg. http://link.springer.com/10.1007/978-3-642-38436-3

Polio, C. (2012). The relevance of second language acquisition theory to the written error correction debate. *Journal of Second Language Writing*, *21*(4), 375–389. https://doi.org/10.1016/j.jslw.2012.09.004

Quené, H., & van den Bergh, H. (2004). On multi-level modeling of data from repeated measures designs: A tutorial. *Speech Communication*, *43*(1–2), 103–121. https://doi.org/10.1016/j.specom.2004.02.004

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A comprehensive grammar of the English language*. Longman. https://doi.org/10.2307/415437

Santos, M., López-Serrano, S., & Manchón, R. (2010). The differential effect of two types of direct written corrective feedback on noticing and uptake: Reformulation vs. error correction. *IJES, International Journal of English Studies*, *10*(1), 131–154. https://doi.org/10.6018/ijes/2010/1/114011

Schmidt, R. (1990). The role of consciousness in second language learning1. *Applied Linguistics*, *11*(2), 129–158. https://doi.org/10.1093/applin/11.2.129

Schmidt, R. (2001). Attention. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 3–32). Cambridge University Press.

Sheen, Y. (2007). The Effect of Focused Written Corrective Feedback and Language Aptitude on ESL Learners' Acquisition of Articles. *TESOL Quarterly*, *41*(2), 255–283. https://doi.org/10.2307/40264353

Sheen, Y. (2010). Differential effects of oral and written corrective feedback in the ESL classroom. *Studies in Second Language Acquisition*, *32*(02), 203–234. https://doi.org/10.1017/S0272263109990507

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

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

Shintani, N., & Ellis, R. (2015). Does language analytical ability mediate the effect of written feedback on grammatical accuracy in second language writing? *System*, *49*, 110–119. https://doi.org/10.1016/j.system.2015.01.006

Shintani, N., Ellis, R., & Suzuki, W. (2014). Effects of Written Feedback and Revision on Learners' Accuracy in Using Two English Grammatical Structures: Effects of Written Feedback and Revision. *Language Learning*, *64*(1), 103–131. https://doi.org/10.1111/lang.12029

Speelman, D. (2014). Logistic regression: A confirmatory technique for comparisons in corpus linguistics. In D. Glynn & R. Justyna A. (Eds.), *Corpus Methods for Semantics: Quantitative studies in polysemy and synonymy* (pp. 487–533). John Benjamins. https://doi.org/10.1075/hcp.43.18spe

Stefanou, C., & Revesz, A. (2015). Direct Written Corrective Feedback, Learner Differences, and the Acquisition of Second Language Article Use for Generic and Specific Plural Reference. *The Modern Language Journal*, *99*(2), 263–282. https://doi.org/10.1111/modl.12212

Stevenson, M., Schoonen, R., & de Glopper, K. (2006). Revising in two languages: A multi-dimensional comparison of online writing revisions in L1 and FL. *Journal of Second Language Writing*, *15*(3), 201–233. https://doi.org/10.1016/j.jslw.2006.06.002

Stiff, R. (1967). The effect upon student composition of particular correction techniques. *Research in the Teaching of English*, *1*(1), 54–75.

Storch, N., & Wigglesworth, G. (2010). Learners' processing, uptake and retention of corrective feedback on writing. *Studies in Second Language Acquisition*, *32*(02), 303–334. https://doi.org/10.1017/S0272263109990532

Swain, M. (1985). Communicative competence: Some roles of comprehensible input and comprehensible output in its development. In S. Gass & C. G. Madden (Eds.), *Input in second language acquisition* (pp. 235–253). Newbury House.

Swain, M. (1995). Three functions of output in second language learning. In G. Cook & B. Seidlhofer (Eds.), *Principle and practice in applied linguistics: Studies in honour of H. G Widdowson.* (pp. 125–144). Oxford University Press. https://doi.org/10.1075/fol.3.2.14fan

Taber, K. (2017). The use of Cronbach's alpha when developing and reporting research instruments in science education. Research in Science Education. Published online June 7, 2017. https://doi.org/10.1007/s11165-016-9602-2

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

Truscott, J. (2001). Selecting errors for selective error correction. *Concentric: Studies in English Literature and Linguistics*, *27*(2), 93–108.

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

Valclav, B. (2018) Statistics in Corpus Linguistics: A Practical Guide, Cambridge, Cambridge University Press, 2018, xix+296 pp., ISBN: 978-1107565241

van Beuningen, C. (2010). Corrective feedback in L2 writing: Theoretical perspectives, empirical insights, and future directions. *International Journal of English Studies*, *10*(2), 1–27. https://doi.org/10.6018/ijes/2010/2/119171

van Beuningen, C., De Jong, N. H., & Kuiken, F. (2012). Evidence on the Effectiveness of Comprehensive Error Correction in Second Language Writing: Effectiveness of Comprehensive CF. *Language Learning*, *62*(1), 1–41. https://doi.org/10.1111/j.1467-9922.2011.00674.x

Van Steendam, E., Rijlaarsdam, G., Sercu, L., & Van den Bergh, H. (2010). The effect of instruction type and dyadic or individual emulation on the quality of higher-order peer feedback in EFL. *Learning and Instruction*, *20*(4), 316–327. https://doi.org/10.1016/j.learninstruc.2009.08.009

Van Steendam, E., Rijlaarsdam, G., & Van den Bergh, H. (2018, June). The effect of observational learning on EFL writers' revision process and product. Presented at the Anéla Applied Linguistics Conference, Egmond aan Zee, the Netherlands.

Vyatkina, N. (2010). The effectiveness of written corrective feedback in teaching beginning German. *Foreign Language Annals*, *43*(4), 671–689. https://doi.org/10.1111/j.1944-9720.2010.01108.x

Wallace, D., & Hayes, J. R. (1991). Redefining Revision for Freshmen. *Research in the Teaching of English*, *25*(1), 54–66.

Wallace, D., Hayes, J. R., Hatch, J. A., Miller, W., Moser, G., & Silk, C. M. (1996). Better revision in eight minutes? Prompting first-year college writers to revise globally. *Journal of Educational Psychology*, *88*(4), 682–688. https://doi.org/10.1037/0022-0663.88.4.682

Appendix A. Sample errors per error category

Subject-verb agreement	People is worried about what they eat.
Article	You will need to have doctor.
Verb	She was send resumes everywhere.
Pronoun	Because of this problems some people think that
Preposition	She wanted to park near to her workplace.
Word form	They are very importants for getting a job.
Subject deletion	Looking back, was a nice person.
Subject repetition	Why is it eating healthy so important?
Sentence structure	The interviewer asked me what was my strength.
Sentence fragment	For example, last year when I went to a nutritionist.
Spelling	Firs of all, you have to be careful when you go on a
Punctuation	Dieting should be done with experts, they know
Capitalization	We have always wanted to visit peru.

Appendix B. Error saliency in learners' initial output

First Impressions			Health and Nutrition		
Error	n	%	Error	n	%
PUNCT	540	31,0	PUNCT	506	32,0
SP	304	17,5	SP	221	14,0
WF	231	13,3	WF	262	16,6
PREP	130	7,5	PREP	123	7,8
SS	93	5,3	SS	112	7,1
SD	79	4,5	CAP	83	5,3
SV	78	4,5	SD	64	4,1
PRON	77	4,4	SV	63	4,0
ART	71	4,1	PRON	59	3,7
CAP	62	3,6	ART	37	2,3
VERB	32	1,8	VERB	25	1,6
FRAG	24	1,4	FRAG	15	0,9
SR	19	1,1	SR	9	0,6
Total	1740	100	Total	1579	100

Note. WF = word form, PREP = preposition, SS = subject repetition, SD = subject deletion, SV = subject-verb agreement, PRON = pronoun, ART = article, VERB = verb, FRAG = fragment, SR = subject repetition, PUNCT = punctuation, SP = spelling, CAP = capitalization