Open access • Journal Article • DOI:10.1111/LANG.12295

# The Differential Effects of Comprehensive Feedback Forms in the Second Language Writing Class — Source link

Marisela Bonilla López, Marisela Bonilla López, Elke Van Steendam, Dirk Speelman ...+1 more authors

**Institutions:** University of Costa Rica, Katholieke Universiteit Leuven

**Published on:** 01 Sep 2018 - Language Learning (John Wiley & Sons, Ltd)

**Topics:** Second language writing and Corrective feedback

# The Differential Effects of Comprehensive Feedback Forms in the L2 Writing Class

Marisela Bonilla López[ab], Elke Van Steendam[a], Dirk Speelman[a], and Kris Buyse[a]
[a]KU Leuven and [b]Universidad de Costa Rica

This study investigated the potential of comprehensive corrective feedback forms as editing and learning tools as well as their effect on learners' cognitive and attitudinal engagement. Low intermediate EFL writers ($N = 139$) were randomly assigned to four experimental conditions (direct corrections on grammatical errors, metalinguistic codes on grammatical errors, direct corrections on grammatical and non-grammatical errors, or metalinguistic codes on grammatical and non-grammatical errors) and a control group (self-correction). Main results from mixed-effect linear models showed that although direct corrections and codes were effective to enhance learners' immediate grammatical and non-grammatical accuracy (i.e., during text revision), a long-term advantage (i.e., four weeks after feedback provision) was only evident for direct corrections. A mental-effort based measure of cognitive load revealed that learners' cognitive load estimates proved significantly lower processing direct corrections targeting grammatical issues. Also, questionnaire answers yielded a significant attitudinal difference between the direct groups and their metalinguistic counterparts.

*Key words.* attitudinal engagement, cognitive load, comprehensive corrective feedback, direct corrective feedback, grammatical and non-grammatical accuracy, metalinguistic corrective feedback.

## 1. Introduction

Over the last years, there has been a growing concern about a mismatch between second language acquisition (SLA) and second language (L2) writing (e.g., Ferris, 2010). Specifically, regarding written corrective feedback (CF) studies, the lack of connection between SLA research findings and their applicability to the L2 writing classroom has prompted some researchers to make a call for an L2-writing interface (e.g., Ortega, 2012). Others, on the other hand, question to what extent SLA research findings have percolated into EFL teachers' feedback practices and stress the need for more classroom-based studies so that teachers do not transfer "findings from previous research that is … remote from classroom realities" (Lee, 2013, p.117). This issue can be best illustrated with those studies that have refuted in a compelling manner any (lingering) arguments against error correction but whose findings may not be representative of some L2 writing contexts because they either excluded revision from their design (e.g., Bitchener & Knoch, 2010a), focused on a few linguistic categories (e.g., Shintani & Ellis, 2013), or took place in a non-L2 writing environment (e.g., van Beuningen, De Jong, & Kuiken, 2012). Consequently, considering the

We are greatly indebted to the students and their instructors for their invaluable contribution to this study. We also wish to thank the anonymous reviewers and the editor, Dr. Pavel Trofimovich, for their insightful comments and constructive advice on earlier versions of this work.
Correspondence concerning this article should be addressed to Marisela Bonilla López, Universidad de Costa Rica, Sede Rodrigo Facio Brenes, Montes de Oca, San José, Código Postal 2060, San José, Costa Rica. Email: marisela.bonilla@ucr.ac.cr

importance of the context that surrounds the feedback, Ferris (2010) suggests one way to address applicability issues: adopting a "blended design" (p.195). According to Ferris, such design adopts the “starting points” of both L2 writing and SLA feedback studies: it not only examines the changes from an initial text to its revision to explore the short-term effect of CF on learners’ immediate accuracy (i.e., the L2 writing starting point) but also incorporates newly produced texts over time to look into the long-term effect of CF on L2 development (i.e., the SLA starting point).

Yet another way to address the need for more applicable findings is by examining a feedback scope that does not conflict with error correction practices of some L2 writing contexts. That is, suggested error correction practices tend to favor corrections on one or a very limited number of error categories (i.e., selective or focused CF), but in L2 writing classrooms, comprehensive CF (i.e., unfocused CF), which targets all or a large number of error categories, is commonly used (Ellis et al., 2008). The crux of the matter is that pedagogical advice cannot be followed without reference to what learners in particular settings need (Hedgcock & Lefkowitz, 1994). Even so, comprehensive CF is said to overwhelm learners, impose a cognitive overload, and hinder their ability to process corrections (e.g., Sheen et al., 2009). Therefore, against this background and in line with Ferris (2010) and Lee (2013), the research base on written CF could benefit from a blended design study that attempts to empirically address accuracy-, cognitive- and attitudinal-related claims about comprehensive CF, which thus far remain under-researched. To this end, the present study provides EFL writers with different forms of comprehensive CF (i.e., CF on all grammatical errors or on both grammatical and non-grammatical errors via direct corrections or metalinguistic codes) to test their differential effect on learners’ immediate and long-term grammatical and non-grammatical accuracy, self-perceived cognitive load, and attitudinal engagement.

## 2. Literature Background

### *2.1. Blended design studies*

Although research on written CF has been steady for over 40 years, different empirical interests have generated varied research questions. As Ferris (2010) explains, “the distinct starting points of L2 writing and SLA research on written CF may cause scholars and practitioners to diverge first in research methodology (and interpretation of resultant findings) and later in application” (p. 191). Therefore, in an attempt to reconcile both lines of work and to “learn from each other and build on one another’s work” (Ferris, 2010, p.191), a study with a blended design delves into the value of written CF both as an instructional intervention to help learners successfully edit their texts and improve their writing and as a learning tool to promote long-term L2 development.

Nevertheless, despite including revision in their design, not all blended design studies report findings from the L2 writing standpoint. For example, Diab (2015) concurs with Polio (2012) in that feedback is useless if learners are not required to do something with it. Therefore, 57 ESL learners, who were assigned to two experimental groups (direct CF plus metalinguistic CF or metalinguistic CF only), had a chance to correct their errors after having received CF. Her findings showed a significant decrease of pronoun errors in the immediate

posttest and of lexical errors in the delayed posttest for learners who received direct plus metalinguistic CF. Notwithstanding the significant evidence of the role of CF as learning tool, Diab (2015) did not report on revision, which remains relevant from an L2 writing stance.

Other studies with a blended design have reported results for both revision and new writings (e.g., van Beuningen et al., 2012), yet only a scant number with marked design and methodological differences has been carried out in L2 composition settings (see Bitchener & Ferris, 2012). For instance, in Ferris (2006) the sample consisted of 92 ESL students, and the treatment involved feedback with codes that targeted 15 error categories. The results showed that learners successfully corrected marked errors in about 80% of the cases. Concerning the long-term effect, the participants significantly reduced the total error ratios from the initial text to the last one. Then, in a study with 31 music majors in ESL writing classes, Chandler (2003) asked participants to write five autobiographical assignments over the course of a semester. The study had an experimental group that received CF (on lexical and grammatical errors) in the form of underlining and had to correct the errors before submitting the next paper. Conversely, learners in the control group, whose errors were also underlined, did the corrections at the end of the semester. Chandler (2003) found that learners in the experimental group outperformed the control group and were able to significantly reduce the total error ratio. However, as pointed out repeatedly in the literature (e.g., Guénette, 2007), in the same way that the lack of control group in Ferris (2006) does not allow to make feedback/no feedback comparisons, neither does the type of control group in Chandler's (2003). Recently, Bonilla, Van Steendam, and Buyse (2017) conducted a study with 52 low- and 39 high-proficiency English and English Teaching majors, who on two different occasions were asked to revise a text after having received comprehensive CF with either direct CF or metalinguistic rule reminders. Also, they wrote a new text three weeks after feedback provision. The results showed no statistically significant differences for proficiency level or interaction between condition and level. They did indicate that a main effect for condition existed: both learner groups in the two experimental conditions were able to correct significantly more grammatical errors during text revision than those who received no feedback and to retain the grammatical accuracy in the delayed posttest. Still, the blended design research base within an L2 classroom setting is limited, and more attention is warranted to comprehensive CF within such context. In this respect, the novelty in our study lies in its examination of various comprehensive CF forms to determine their editing and L2 learning potential (or lack thereof) in the L2 writing class.

### *2.2. The "right" amount of written CF*

Research on comprehensive CF merits further investigation for pedagogical and theoretical reasons. First, one of the (many) pedagogical decisions that L2 teachers are confronted with is how much written CF they should provide. For this reason, in instructional contexts where comprehensive CF is called for, correcting a few linguistic categories could be hard to implement. For instance, a teacher's goal may be to train learners to "produce high-quality final products" (Bitchener & Ferris, 2012, p.117) or to edit an entire text to improve overall linguistic accuracy (Hartshorn et al., 2010). Thus, in such cases, correcting

all or a large array of error categories may be a pedagogical need rather than a choice. Second, although studies to date have proved that written CF can be a useful text revision tool (e.g., Ferris & Roberts, 2001), which may also lead to L2 learning (e.g., Bonilla et al., 2017), the body of research remains insufficient to settle things pertaining to the amount of feedback to be provided (Bitchener & Storch, 2016). In this respect, out of studies that have addressed the effect of comprehensive/unfocused CF, we make a distinction between those that have incorporated a focused/selective group in their design as a baseline for comparison and those that have not.

First, from studies that have included a baseline comparison with a focused/selective CF treatment, conflicting findings have emerged. Also, the extent to which some of these studies answer questions about unfocused/comprehensive CF is debatable[1]. For example, Ellis et al. (2008) randomly assigned 49 EFL intermediate learners to three conditions: direct focused CF on (definite and indefinite) article errors; direct unfocused CF on article errors and other linguistic categories; and no feedback. Their results revealed that both experimental groups had accuracy gains from pre-test to post-test as opposed to the control group (whose grammatical accuracy declined), yet the differences between the direct focused group and the direct unfocused group did not reach statistical significance. This prompted Ellis et al. (2008) to conclude that focused and unfocused CF were equally effective. A year later, Sheen et al. (2009) conducted a study with 80 intermediate ESL students, who received (1) direct focused CF on article errors, (2) direct unfocused CF on article, copula "be", past tense, and preposition errors, (3) writing practice, or (4) no feedback at all. They found that concerning the accuracy of use of the English article, group 1 significantly outperformed groups 2, 3, and 4 in the immediate posttest. Also, in the delayed posttest, group 1 performed better than group 4. Hence, the researchers affirmed that "focused CF is more effective than unfocused CF" (Sheen et al., 2009). More recently, Frear and Chiu (2015) investigated the changes over time of both learners' use of weak verbs (regular past tense verbs) and their overall grammatical accuracy. The study with 42 Chinese EFL learners had two experimental conditions (focused indirect CF [on weak verbs] and unfocused indirect CF [on all errors]) and a control group (without CF provision). The authors found that learners in the experimental groups significantly outperformed the control group in both posttests, yet no significant differences were found between conditions for accurate use of weak verbs or for total accuracy. The results showed accuracy improvement from pretest to posttest but no continued improvement from immediate posttest to delayed posttest, which the researchers concluded could be due to the number of feedback sessions (one), the indirect nature of the treatment, and the amenability to correction of the targeted linguistic features.

Second, from studies that have investigated unfocused/comprehensive CF alone, firm answers remain difficult to obtain due to research design or comparability issues. For instance, some studies have been criticized for their type of control group (e.g., Sheppard, 1992) or the presence of a dissimilar incentive across conditions (e.g., Semke, 1984). Others have been able to provide clearer evidence about the effects of comprehensive CF while at the same time addressing research design issues of previous studies. For instance, Truscott and Hsu (2008) assigned 47 EFL graduate students to an experimental group receiving CF

with underlining and the control group receiving no feedback. All students were required to revise a picture-based narrative story they had written a week before. Then, a week later students wrote a new narrative. The results showed that the experimental group reduced the error rate significantly more than the control group, yet that grammatical accuracy improvement was not significantly sustained a week later in a new text, prompting the researchers to claim that CF has no value as a teaching device. However, concerning the feedback effect as a learning tool, the findings differ from three other studies that did find evidence of learning as a result of comprehensive CF. For example, van Beuningen et al. (2008, 2012) showed that ESL learners in a biology class could successfully process feedback that targeted a large array of errors. Their evidence demonstrated that both comprehensive direct and metalinguistic CF with codes helped EFL learners to significantly enhance a revised text. As for the long term-effect of CF, in their first study only direct CF had a sustained effect one week after feedback provision (van Beuningen et al., 2008), whereas in the study that followed, both direct and coded CF groups maintained the accuracy gains four weeks after treatment (van Beuningen et al, 2012). Then, with similar revision and learning benefits, Bonilla et al. (2017) demonstrated that the grammatical accuracy of EFL university writers with a low- and high-proficiency level could benefit from comprehensive CF with direct corrections or metalinguistic rule reminders. Clearly, Truscott and Hsu (2008), van Beuningen et al. (2008, 2012), and Bonilla et al. (2017) are not comparable, and differing results may have been due to differences in the longitudinal period examined, the number (and type) of targeted linguistic features, and the feedback strategies employed, which differed in degree of explicitness.

To this day then, a lingering gap in our current knowledge of comprehensive CF is how much comprehensive correction of grammatical and non-grammatical errors learners are able to handle or not. Arguably, accuracy (both grammatical and non-grammatical) constitutes one important component that speaks of text quality in L2 academic contexts (Hyland, 2003), yet only one comprehensive CF study has looked into two accuracy types (i.e., van Beuningen et al., 2012). With this in mind, we incorporated a baseline for a comparison in which learners receive different forms of comprehensive CF. Such research design could shed some light on issues concerning comprehensive CF which thus far have not been empirically tested. For instance, if comprehensive CF is "ineffective" (Ellis et al., 2008, p.368), can comprehensive attention to grammatical issues alone or to both grammatical and non-grammatical issues (i.e., spelling punctuation, and capitalization) hinder learners' immediate and long-term accuracy improvement? Also, could there be a differential effect on learners' cognitive or affective response as a result of processing comprehensive CF forms?

### *2.3. The cognitive and affective load of comprehensive CF*

Models of L2 acquisition such as those by Robinson (1995, 2005) and Schmidt (1990, 2001) have provided theoretical ground for claims that pertain to the effects of comprehensive CF and leaners' ability to process it (Bitchener, 2012). For instance, Robinson's (2003, 2005) model, also known as The Cognition Hypothesis, distinguishes between differences in the processing demands of tasks and the resources that learners bring to perform such tasks. He claims that tasks vary in the demands they impose on learners'

attention (Robinson, 2003) and defines task demands as "the attentional, memory and reasoning demands…that increase the mental workload the learner engages in performing the task" (Robinson, 2001, p.302). Similarly, task demands play an important role in Schmidt's (1990) noticing hypothesis because they greatly determine what learners are able to notice. This in turn is relevant for subsequent language acquisition given that noticing is the "conscious attention to the form of input" (Robinson, 1995, p.284). Hence, if a task demands more than learners' cognitive abilities can handle, it follows that their chances of consciously noticing the input and subsequently internalizing it could be reduced. Other researchers have honed in on learners' attentional capacity (e.g., Skehan, 1998), but unlike Robinson (1995 and elsewhere), they have posited that attention is capacity-limited and are less positive about the likely effects on performance as a result of such capacity constraints. For example, applied to written CF, Skehan's view can imply that limitations in working memory would cause competition between attention to grammatical issues and attention to non-grammatical issues to such a degree that either type of accuracy would deteriorate. Conversely, drawing on Robinson's, the increasing complexity of tasks may "push for greater accuracy and L2 production" (Robinson & Gilabert, 2007, p.162) so much so that attention to comprehensive CF forms may not necessarily be detrimental and both grammatical and non-grammatical accuracy could still be achieved.

Even so, comprehensive CF is thought to impose more attentional demands than learners' attentional capacity can allocate. Cognitive-related claims state, for example, that directing learners' attentional resources to a broad range of issues could place a heavy cognitive load (Bitchener, 2008) and tax learners' ability to process the feedback (Sheen, 2007). These effects have to do with what Bitchener (2008) describes as "the difficulty that ESL learners experience in trying to cope with information overload" (p. 109). Similar to Bitchener (2008), Ellis et al. (2008) believe that "[a] mass of corrections directed at a diverse set of linguistic phenomena (and perhaps also at content and organizational issues) is hardly likely to foster the noticing and cognizing that may be needed for CF to work for acquisition" (p.368). In other words, comprehensive CF may not bring about L2 acquisition because noticing is not likely to occur. Also, concerning cognitive processing and along the lines of Ellis et al. (2008), Evans et al. (2010) argue that learners may not benefit from comprehensive written CF when as a result of feeling overwhelmed, they neither process nor learn from the feedback. Thus, comprehensive CF is considered "overloading" (Sheen et al., 2009, p.559). Nevertheless, there is an empirical void to substantiate such a claim. The bottom line is that while the volume of work measuring the cognitive load of complex cognitive tasks is considerably large in educational research domains such as physics (Sweller, 1988), computer programming (Paas & Van Merriënboer, 1994b), and mathematics (Paas, 1992) (for a summary, see Paas & Van Merriënboer, 1994a; Sweller, Ayres, & Kalyuga, 2011), no study in applied linguistics, to be the best of our knowledge, has sought to measure the cognitive load of tasks that involve revision after differing written CF scopes.

In addition, learners' affective response to written CF has prompted some researchers to advise L2 teachers to shy away from grammar correction generally (e.g., Semke, 1984) and comprehensive CF particularly (e.g., Truscott, 2001). To illustrate, when Truscott (1996) built his case against grammar correction, he did so by referring to the side effects it has on learners' grammatical accuracy and "students' attitudes" (p.328). Interestingly, it seems that

the debate he ignited was enough to concentrate most research efforts on the former but not so much on the latter. To date, our knowledge of learners' attitudes towards written CF is largely based on descriptive studies (e.g., Incecay & Dollar, 2011). Although "written feedback is more than marks on a page" (K. Hyland & Hyland, 2006, p.84), few experimental feedback studies (e.g., Diab, 2015) have taken into consideration affective variables—such as learners' attitudinal engagement[2] with specific treatment—to further understand feedback outcomes. Consequently, whether or not L2 learners feel that comprehensive CF is an "unpleasant" (Truscott, 1996, p.352), "overwhelming" (Bitchener & Ferris, 2012, p.117), "confusing" (Sheen et al., 2009, p.567), "discouraging" (Truscott, 2001, p.93) , or "de-motivating" (Bitchener & Ferris, 2012, p. 128) practice is in much need of further scrutiny. If indeed various forms of comprehensive CF are overburdening, on the one hand, and unwelcome, on the other hand, the theoretical and practical repercussions would be worth noting given the evidence that indicates that a high cognitive load (cf. Paas & Van Merriënboer, 1994a) and feedback resistance (cf. Storch & Wigglesworth, 2010) could impede learning.

## 3. The current study

The present study aimed to expand the research base on written CF by investigating the effect that different comprehensive CF forms (i.e., CF on grammatical errors or on both grammatical and non-grammatical errors provided directly or with metalinguistic codes) have on EFL writers' immediate and long-term grammatical and non-grammatical accuracy, self-perceived cognitive load, and attitudinal engagement. Specifically, the research questions (RQ) that guided the study add to current understanding of comprehensive CF by addressing (accuracy, cognitive, and attitudinal) under-researched issues previously identified in the review of literature.

First, bearing in mind the need for empirical evidence that applies to L2 writing contexts and that conforms with error correction practices other than selective CF, this study examined the value of comprehensive CF as an editing (i.e., the L2 writing perspective) and learning tool (i.e., the SLA perspective) in the L2 writing class:

RQ1. To what extent do comprehensive feedback forms lead to improved (grammatical and non-grammatical) accuracy during text revision and in new writings over time?

In this respect, previous accuracy-related claims about comprehensive CF (e.g., Ellis et al., 2008) could imply that L2 learners may not be able to succeed at processing a large number of corrections. Nonetheless, considering the principle of Transfer Appropriate Processing (TAP), a match between the testing condition and the learning outcome should occur (Lightbown, 2008). Thus, two plausible hypotheses (H) are the following:

H1. Learners who attend to corrections on grammatical errors will show immediate and long-term gains in grammatical accuracy.

H2. Learners who attend to corrections on both grammatical and non-grammatical errors may show immediate and long-term gains in grammatical and non-grammatical accuracy.

Second, because virtually no feedback study has been conducted on learners' cognitive load after revision with comprehensive CF, the research question below was deemed desirable:

RQ2. What is learners' self-perceived cognitive load after revision with comprehensive CF forms?

More specifically, we could assume that learners' cognitive response to comprehensive CF will hinge upon the treatment they receive. That is, if we bear in mind (a) the number of targeted features, (b) the explicitness of the corrective information, and (c) the level of engagement that the revision task requires, some feedback processing forms in this study may be more cognitively complex (e.g., text revision after metalinguistic CF with codes on all grammatical and non-grammatical issues issues) than others (e.g., text revision after direct corrections on grammatical issues). Therefore, if comprehensive CF is overloading (Sheen et al., 2009), more cognitively complex feedback processing forms will impose a higher self-perceived cognitive load than less cognitively complex ones (H3).

Finally, we explored attitudinal-related claims to address the affective response that comprehensive CF is thought to generate from L2 learners. Hence, given the scarce findings available from feedback studies about this issue, we asked the following:
RQ3. What (if any) is the effect that comprehensive feedback forms have on learners' attitudinal engagement?

Similar to learners' cognitive response, if comprehensive CF is unwelcome (cf. Truscott, 2001), more cognitively complex feedback processing forms will render a less favorable attitudinal response than less cognitively complex ones (H4).

## 4. Methods

### *4.1. Participants and instructional context*

This study took place in the main campus of an urban public university in Costa Rica, which comprises students from both urban and rural areas. More specifically, the present study was carried out in the School of Modern Languages with 139 participants (53 male and 86 female, mean age = 21, *SD* = 4.11) majoring in English (n = 102) or English Teaching (n = 37), which implies that all participants were pursuing a career as English professionals for which mastery of the target language is primordial. They were enrolled in an integrated English course that not only teaches basic writing conventions but also uses writing as a vehicle to teach the target language (cf. Manchón, 2011). Students in this course met four days a week, three hours a day. Their native language was Spanish, and their mean English proficiency level was lower intermediate (*SD* = .79) as ascertained by Oxfords' Quick Placement test (QPT) (see QPT results in section 4.5.1.). The participants were randomly assigned to five groups: direct CF on grammatical errors (n = 29), metalinguistic CF with codes on grammatical errors (n = 28), direct CF on grammatical and non-grammatical errors (n = 27), metalinguistic CF with codes on grammatical and non-grammatical errors (n = 28), and a control group (n = 27).

### *4.2. Feedback strategies*

Two of the most examined feedback techniques in the CF research base (see Kang & Han, 2015; Liu & Brown, 2015 for a meta-analysis) are direct CF and metalinguistic CF with codes—the latter also referred to in the literature as coded (e.g., Sampson, 2012) or indirect CF (e.g., van Beuningen et al., 2012). Specifically, in this study we employed the aforementioned feedback types bearing in mind (1) that they are two commonly used

strategies within the instructional context of our investigation and (2) that despite the large research base on CF, applicable or clear-cut findings are hard to obtain due to either marked differences in research design (cf. section 2.1.) or research methodological issues (cf. section 2.2.).

*4.3. Treatment and control*

The study had four experimental groups. Learners received either direct CF on grammatical errors (hereafter DCF+G), metalinguistic CF with codes on grammatical errors (hereafter ME+G), direct CF on grammatical errors and non-grammatical errors (hereafter DCF+GN), and metalinguistic CF with codes on grammatical errors and non-grammatical errors (hereafter ME+GN). Specifically, grammatical correction (i.e., morphology and syntax) targeted all linguistic errors, for example, errors in word form (e.g., singular/plural), word order (e.g., sentence structure), agreement (e.g., pronoun), incomplete sentences (e.g., fragment), and unnecessary insertion or faulty omission of elements. Non-grammatical correction targeted orthographical errors: spelling, punctuation, and capitalization. The study also had a self-correction (i.e., control) group (hereafter SC), where learners did not receive any corrections in their text. This means that the suppliance of feedback in the experimental groups was other-provided (i.e., the researcher) whereas in the control group, it was self-provided (i.e., the learners themselves). We operationalized DCF and ME as defined in Ellis (2009b) and Bitchener and Storch (2016). Thus, the former consisted of providing learners with the correct target language form above the error, whereas the latter entailed underlining the error and providing a code above to know its type. The coding system consisted of 14 different categories of error types, which were then classified into two: grammatical and non-grammatical. For each code that was used in a learner's text, the code and its spelled out form were written at the bottom of the composition (see Appendix S1 in the Supporting Information for a sample coding system). The reason for spelling out the codes was contextual: although learners were already familiar with the coding system as it was used in the previous course, we felt the need to refresh their memory after a two-month vacation period.

*4.4. Design and procedures*

The entire data collection process took six weeks during which five sessions were carried out (see Figure 1). In session 1 of week 1, the participants took a proficiency test and wrote the initial text (pretest). For writing the pretest, learners had 30 minutes. Two days later in session 2, the students were allotted 15 minutes to study a copy of their initial text, which had been corrected (or not) according to the condition they had been assigned to (see Appendix S2 in the Supporting Information for studying instructions). Hence, learners in the experimental groups studied the feedback provided whereas those in the control group studied the text for self-correction purposes. Once they finished, the copies were taken away, and students were asked to take 30 minutes to revise the same text while looking at their original (uncorrected) writing piece instead (see Appendix S3 in the Supporting Information for revision instructions).

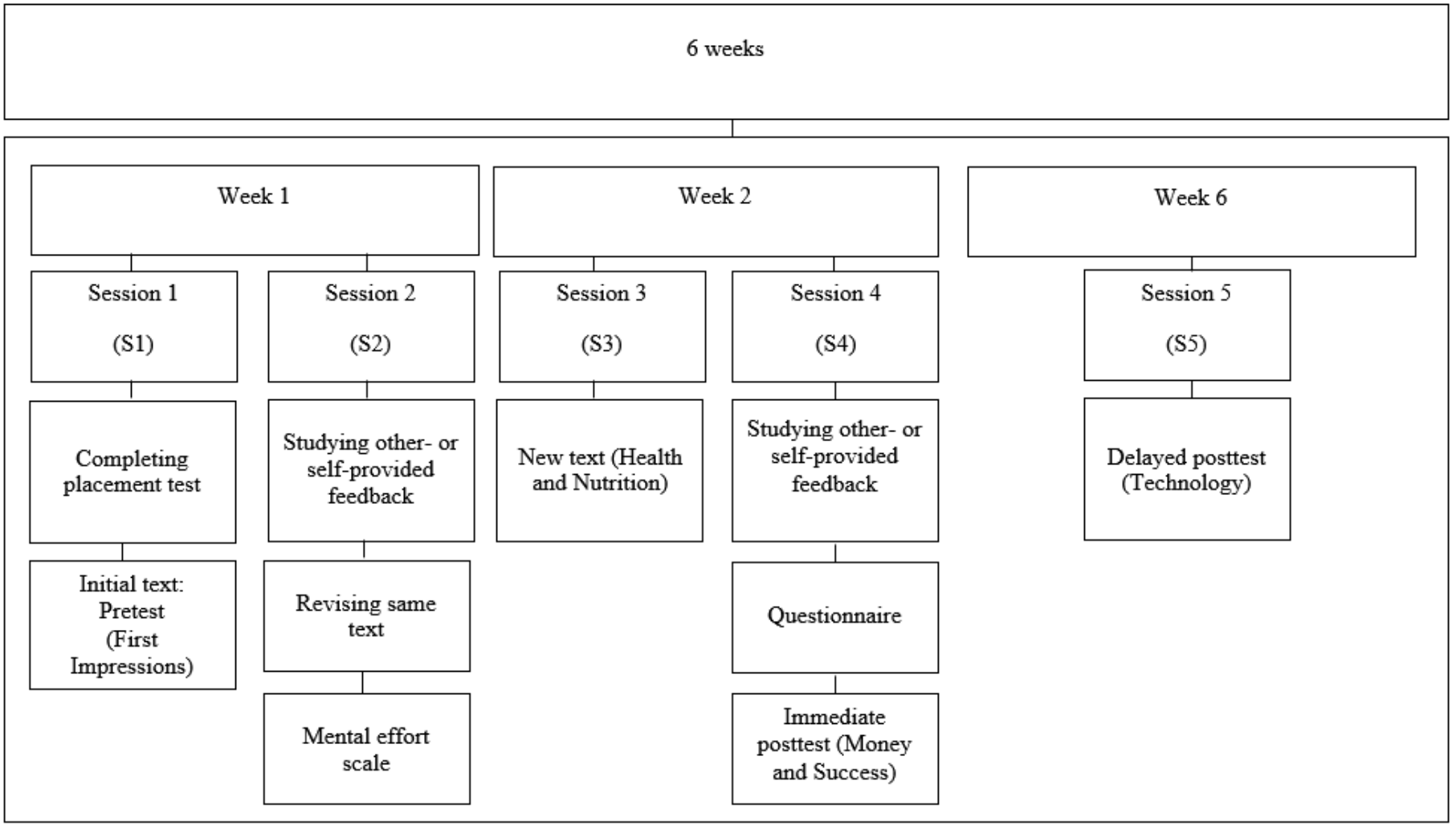



**Figure 1.** Study design

Our rationale behind this revision procedure was pedagogically motivated: we did not want the revision task to be a copying exercise. In fact, we concur with (Polio, 2012) in that revising a text while looking over the corrections is "[f]rom both a pedagogical and a theoretical perspective… the least interesting" (p.377) way to examine a writer. Besides, it is worth noting that for studying the copy with other- (in the experimental groups) and self-provided (in the control group) feedback, task instructions did not hint in any way at the possibility that there was going to be a revision session afterwards or that the copy was going to be taken away (cf. Appendix S2). Students' attention was then drawn to the task at hand to reduce the chances of memorization as it has previously occurred in three-stage writing task studies where task instructions (e.g., knowing that the feedback will not be available for a subsequent revision task) played a role in learners' attentiveness to feedback and strategies to process it (e.g., Santos, López-Serrano, & Manchón, 2010)[3]. The time on task for both studying the copy of the composition and revising the text was the same in all conditions, and it was decided upon standard practice in the instructional setting of this study. We highlight that we took away the copies of learners in the control group because they jotted down corrections while self-providing feedback. Therefore, allowing them to keep those notes would have allowed them to copy their self-provided corrections, adding an undesirable variable in our study. After text revision, learners completed the mental effort scale. Then, a week later in session 3, students were given 30 minutes to write a new text, which they studied two days later in session 4 according to their assigned condition and following the same procedures as in session 2. The time on task and the feedback conditions did not change. Later, after having finished studying their text and having had the copy of their compositions taken away as in session 2, learners in the experimental groups answered the questionnaire of attitudinal engagement (see section 4.5.3.) to express their reactions toward the treatment. The time allotted for completing the questionnaire was ten minutes. After that, all learners wrote a new piece (immediate posttest) for which they had 30 minutes. Four weeks later, in session 5 students had 30 minutes to write a new text (delayed posttest). Finally, all learners were informed from the start of the study that because the topics and tasks were part of the curriculum, the texts could eventually become drafts of future graded compositions at the end of the study if deemed desirable by their instructor. This decision was made weeks prior to the data collection when both the researchers and the instructors met; the underlying reason was to ensure a well-balanced situation for both parts: the teachers would compensate for some of their class work and the researchers would prevent absenteeism.

*4.5. Materials*

*4.5.1. Placement test*

We administered Oxford's Quick Placement Test (QPT) to better ascertain learners' English proficiency level. The results indicated that the mean English proficiency level was lower intermediate (*SD* = .79). Others were elementary (n = 12), upper intermediate (n = 17), advanced (n = 15), and very advanced (n = 1). However, keeping in mind the random assignment of participants to conditions, significant differences in proficiency level were unlikely and were not found (proficiency level [$F(4,139) = 1.864$, $p.= .120$, $\eta_p^2 = .05$]).

*4.5.2. Writing tasks*

To measure the feedback effect on grammatical and non-grammatical accuracy, learners produced four texts, received feedback on two of them, and revised one. They were asked to write a 175-word opinion composition about chapter-related topics, which means that all learners were exposed to the thematic vocabulary during class activities. The chapter-related topics were the following (cf. Figure 1): First Impressions ("Do you agree that people should exaggerate the truth or outright lie in their resume if that will help them to get a job? Why or why not?"), Health and Nutrition ("Do you agree that people should diet more and eat less to live a healthy life? Why or why not?"), Money and Success ("Do you agree that success equals big money? Why or why not?"), and Technology ("Do you agree that depending on technology contributes to losing control of our lives? Why or why not?"). The instructions in all tasks consisted of a question that elicited their opinion about a given topic and the same prompt to elaborate on it (see Appendix S4 in Supporting Information for a sample writing task). Similar to other feedback studies (e.g., Lavolette, Polio, & Kahng, 2015; van Beuningen et al., 2012), the writings tasks were not counterbalanced. The reason was the fact that all writing tasks and topics were part of the curriculum, which (besides bolstering the ecological validity of the study) implied that they had to be administered in accordance with the course outline.

*4.5.3. Questionnaire of attitudinal engagement*

Given that comprehensive CF is thought to overwhelm and demotivate L2 learners, we probed learners' affective response, namely their attitudinal engagement with comprehensive CF forms. To this end, we adapted the Questionnaire of Attitudinal Engagement and Feedback Preferences in Bonilla et al. (2017). Hence, only students in the experimental groups (n = 112) answered the instrument. Reactions to self-provided feedback were not considered in the analyses. The adapted questionnaire had acceptable internal consistency ($\alpha$ =.73), and it was administered immediately after the last feedback session to shorten the reference period and avoid biased feedback estimates (De Leeuw & Dillman, 2008). The questionnaire consisted of 10 scale items where learners had to indicate from 1 (not at all) to 5 (totally) to what extent each statement best described their attitudes towards the feedback.

*4.5.4. Mental effort scale*

To measure the cognitive load of comprehensive CF forms, we administered Paas's (1992) mental-effort based scale[4]. Answers about why "measures of mental effort constitute the essence and the best estimator of cognitive load" (Paas & Van Merriënboer, 1994a, p.357) can be obtained in Paas and Van Merriënboer's schematic representation of cognitive load. The authors define cognitive load as "a multidimensional construct that represents the load that a particular task imposes on the cognitive system of a learner" (p.353) and in their schematic representation of the construct, they explain that mental effort is believed to be an actual representation of cognitive load because it is "the aspect of cognitive load that refers to the cognitive capacity that is actually allocated to accommodate the demands imposed by the task" (Paas, Tuovinen, Tabbers, & Van Gerven, 2003, p.64). Specifically, prior to Paas's scale there was not a subjective measure of cognitive load (Sweller et al., 2011), which is

why this index has been widely adapted in cognitive load research as an offline technique (i.e., administered after task performance) (Leppink, Paas, Van der Vleuten, Van Gog, & van Merriënboer, 2013). The underlying assumption behind it is that people are capable of reflecting on their cognitive processes and self-rate their perceived intensity of mental effort (Paas et al., 2003). Against this background, we adapted the instructions of Paas's mental-effort based measure of cognitive load, which were originally related to statistics. The instrument had acceptable internal consistency (α =.78) and consisted of a 9-point item scale ranging from 1 (very, very low mental effort) to 9 (very, very high mental effort), where learners reported on their perceived amount of mental effort after, in this case, processing comprehensive CF forms (see Appendix S5 in Supporting Information).

*4.6. Coding and analysis*

For examining the effects of feedback on learners' immediate grammatical accuracy, as in Bonilla et al. (2017), we traced each error and labeled it based on the text revision behavior under study: GEC (i.e., grammatical error successfully corrected) or NGEC (i.e., non-grammatical error successfully corrected). We did not employ an error-words ratio (e.g., Chandler, 2003) because it may not accurately depict learners' enhanced (or not) immediate accuracy in cases in which the absence of errors is due to avoidance and deletion from the text (Murphy & Roca de Larios, 2010) rather than to successful error correction. Therefore, after having traced the errors and labeled learners' text revision behavior, we computed a new variable: the number of errors that were successfully corrected during text revision divided by the total number of errors in the initial text. This was done for grammatical and non-grammatical errors. Pertaining to learners' overall (grammatical and non-grammatical) accuracy in newly produced texts across time (i.e., L2 development), we used an overall accuracy measure. It consisted of the total number of (grammatical or non-grammatical) errors divided by the total number of words multiplied by 10 (Bonilla et al., 2017; van Beuningen et al., 2012).

All pen-and-paper compositions (n = 695) were converted to Word using Dragon Naturally Speaking 11.0. This speech recognition software, which was used merely for transcription purposes, allowed the first researcher and a research assistant to dictate each composition and obtain a verbatim digital version. Then, the first researcher blindly coded all texts for grammatical and non-grammatical accuracy. To determine interrater reliability, three experienced writing teachers from the institution where the study took place recoded 40 texts randomly selected from the immediate posttest. The rationale for recoding 40 texts was contextual: the three teachers already had a full-time work load, and when the recoding time came (i.e., three months later), it was also examination period. This meant that coding 40 texts was the only manageable extra work load they could handle at the time. Therefore, to meet the required 10 percent of coded data, an independent experienced rater coded 70 texts, which were randomly chosen from the five sessions (N = 14 per session). Ten months later, the first researcher recoded 10 percent of the data to establish intrarater reliability. Table 1 shows the Cronbach's alpha scores for two measures. As can be seen, all alphas reached acceptable reliability (Taber, 2017).

**Table 1**
*Alpha Scores for Interrater and Intrarater Reliability*

| | Grammatical accuracy | Non-grammatical accuracy |
|---|---|---|
| Interrater[a] | .834 | .801 |
| Interrater[b] | .905 | .917 |
| Intrarater | .955 | .962 |

[a]Reliability scores from three teachers. [b]Reliability scores from external rater.

Concerning the questionnaire, we subjected the items to an exploratory factor analysis, which yielded three components (see Appendix S6 in Supporting Information for scale items and summary of factor loadings of the exploratory factor analysis). The components grouped items dealing with learners' attitudes towards the feedback regarding its overall usefulness, comprehensibility, and emotional burden. Therefore, we labeled the components *utility*, *comprehensibility*, and *burden*, respectively. The factor analysis with a Varimax (orthogonal) rotation yielded a determinant value of .046, a Kaiser-Meyer Olkin (KMO) measure of .670, and a significant Bartlett's test ($p$ = < .000). After confirmation of internal consistency (cf. section 4.5.3.) and factor analysis loadings, we proceeded to create a composite score of the constructs (i.e., utility [item 1 + 7 + 4 + 5], comprehensibility [item 9 + 3 + 10 + 6], and burden [item 2 + 8]).

Finally, with the obtained mental-effort scale ratings, the computed variables for immediate and overall (grammatical and non-grammatical) accuracy, and the composite score of the three constructs, we proceeded to enter the aforementioned dependent variables in a mixed-effect linear model (also called multi-level models). Using treatment coding, we opted for mixed-effect models because they offer a more versatile and technically more sophisticated alternative to traditional ANOVAs and repeated measures ANOVAs for the analysis of repeated measures and other types of grouped data (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, Bretz, & Westfall, 2008); effect size measures using both $R^2$ and $\Omega^2$ (for mixed-effect linear models) and both $R^2$ and adjusted $R^2$ (for linear models) were calculated by the function r2 from the sjstats package (Ludecke, 2017).

## 5. Results

After presenting the preliminary analyses, this section will report the results pertaining to the effect of comprehensive feedback forms on learners' immediate grammatical and non-grammatical accuracy (section 5.2.), grammatical and non-grammatical development (section 5.3.), self-perceived cognitive load (section 5.4.), and attitudinal engagement (section 5.5.). To this purpose, separate tables that summarize the descriptive statistics for all response variables (Table 2) and the significant post-hoc comparisons (Table 3) will also be provided.

*5.1. Preliminary analyses*

At the outset of this study, we did not find initial differences in English proficiency level, $F(4,139) = 1.864$, $p = .120$, $\eta_p^2 = .05$); overall grammatical accuracy, $F(4,139) = .386$, $p = .818$, $\eta_p^2 = .01$; overall non-grammatical accuracy, $F(4,139) = .711$, $p = .586$, $\eta_p^2 = .02$; or perceived cognitive load, $F(4,139) = 2.086$, $p = .086$, $\eta_p^2 = .05$ (see Appendix S7 in Supporting Information for descriptive statistics).

*5.2.Feedback effect on immediate grammatical and non-grammatical accuracy*

The mixed-effect model revealed a statistically significant main effect for condition on grammatical errors successfully corrected, $F(4,133) = 26.47$, $p < 0.001$, $R^2 = .44$, $R^2$ adjusted = .42 (see Appendix S8 and S12 in Supporting Information for full model and a summary of the significant fixed effects kept in the models, respectively). During text revision, experimental groups DCF+G ($p < 0.001$, $SE = 0.069$), ME+G ($p = 0.009$, $SE = 0.070$), DCF+GN ($p < 0.001$, $SE = 0.070$), and ME+GN ($p < 0.001$, $SE = 0.070$) significantly outperformed the control group. Also, the DCF+G group corrected significantly more grammatical errors than groups ME+G ($p < 0.001$, $SE = 0.069$), DCF+GN ($p = 0.021$, $SE = 0.069$), and ME+GN ($p < 0.002$, $SE = 0.069$). The ME+G group corrected a significantly lower number of grammatical errors than DCF+GN ($p = 0.012$, $SE = 0.070$). The difference between ME+GN and ME+G as well as between ME+GN and DCF+GN did not reach statistical significance (see Tables 2 and 3).

Also, the mixed-effect model yielded a significant main effect for condition on non-grammatical errors successfully corrected, $F(4,133) = 22.07$, $p < 0.001$, $R^2 = .42$, $R^2$ adjusted = .40 (see Appendix S8 and S12 in Supporting Information for full model and a summary of the significant fixed effects kept in the models, respectively). The DCF+GN group did significantly better at correcting non-grammatical errors in revised texts than groups SC ($p < 0.001$, $SE = 0.061$), DCF+G ($p < 0.001$, $SE = 0.060$), and ME+G ($p < 0.001$, $SE = 0.061$). The same was true for ME+GN, which significantly outperformed the SC ($p < 0.001$, $SE = 0.060$), DCF+G ($p < 0.001$, $SE = 0.059$), and ME+G groups ($p < 0.001$, $SE = 0.059$). No significant differences were found between DCF+GN and ME+GN (see Tables 2 and 3).

*5.3.Feedback effect on grammatical and non-grammatical development*

The mixed-effect model revealed a significant interaction effect for condition and time for overall grammatical accuracy, $X^2_4 = 44.31$, $p < 0.001$, $R^2 = .72$, $\Omega^2 = .71$ (see Appendix S9 and S12 in Supporting Information for full model and a summary of the significant fixed effects kept in the models, respectively). Learners in experimental groups DCF+G ($p < 0.001$, $SE = 0.017$), ME+G ($p < 0.001$, $SE = 0.017$), and DCF+GN ($p < 0.001$, $SE = 0.017$) could significantly improve their grammatical accuracy in the long term more than those who received no CF. More significant contrasts were found for the DCF+G group, whose grammatical accuracy gain over time was significantly higher than that of ME+GN ($p = 0.001$, $SE = 0.017$) (see Tables 2 and 3).

**Table 2**

*Descriptive Statistics for Immediate Accuracy, Accuracy Development, Perceived Cognitive Load, and Attitudinal Engagement*

| | DCF+G (n = 29) | | ME+G (n = 28) | | DCF+GN (n = 27) | | ME+GN (n = 28) | | SC (n = 27) | |
|---|---|---|---|---|---|---|---|---|---|---|
| Response variable | *M* | *SD* | *M* | *SD* | *M* | *SD* | *M* | *SD* | *M* | *SD* |
| Immediate accuracy | | | | | | | | | | |
| Grammatical | .749 | .273 | .316 | .307 | .543 | .291 | .505 | .254 | .088 | .165 |
| Non-grammatical | .173 | .232 | .135 | .162 | .474 | .283 | .538 | .279 | .283 | .288 |
| Accuracy development | | | | | | | | | | |
| Grammatical | | | | | | | | | | |
| Session 1 | .387 | .247 | .325 | .237 | .331 | .252 | .353 | .260 | .312 | .261 |
| Session 3 | .331 | .218 | .271 | .189 | .349 | .273 | .344 | .292 | .310 | .218 |
| Session 4 | .183 | .159 | .213 | .200 | .212 | .180 | .323 | .325 | .355 | .209 |
| Session 5 | .127 | .133 | .222 | .141 | .182 | .146 | .290 | .217 | .394 | .203 |
| Non-grammatical | | | | | | | | | | |
| Session 1 | .365 | .204 | .309 | .167 | .400 | .213 | .363 | .228 | .349 | .207 |
| Session 3 | .359 | .240 | .224 | .153 | .355 | .185 | .411 | .248 | .333 | .189 |
| Session 4 | .354 | .250 | .281 | .185 | .228 | .157 | .294 | .217 | .373 | .219 |
| Session 5 | .348 | .183 | .348 | .183 | .191 | .204 | .293 | .189 | .387 | .163 |
| Perceived cognitive load | 4.76 | 1.66 | 5.71 | 1.27 | 5.59 | 1.42 | 6.68 | .612 | 4.48 | 1.39 |
| Attitudinal engagement[a] | | | | | | | | | | |
| Utility | 17.14 | 2.34 | 16.07 | 2.90 | 16.33 | 2.52 | 16.11 | 2.84 | | |
| Burden | 8.79 | 1.59 | 8.14 | 2.20 | 8.15 | 2.16 | 7.54 | 2.11 | | |
| Comprehensibility | 18.10 | 2.17 | 16.36 | 3.08 | 18.48 | 1.94 | 16.39 | 3.17 | | |

*Note.* DCF+G = direct CF on grammatical errors; ME+G = metalinguistic CF with codes on grammatical errors; DCF+GN = direct CF on grammatical and non-grammatical errors; ME+GN = metalinguistic CF with codes on grammatical errors and non-grammatical errors; SC = self-correction with no feedback provided.

[a] Learners in the SC condition were not considered in the analyses.

**Table 3**
*Summary of Significant Post-hoc Comparisons per Response Variable*

| | Comparison | *b* [95% CI] |
|---|---|---|
| Immediate accuracy | | |
| Grammatical | ***DCF+G > SC | 0.67 [0.48, 0.86] |
| | **ME+G > SC | 0.23[0.04, 0.42] |
| | ***DCF+GN > SC | 0.46 [0.26, 0.65] |
| | ***ME+GN > SC | 0.41 [0.22, 0.61] |
| | ***DCF+G > ME+G | 0.44 [0.63, 0.25] |
| | **DCF+G > ME+GN | 0.25 [0.44, 0.06] |
| | *DCF+GN > ME+G | 0.22 [0.03, 0.42] |
| | *DCF+G > DCF+GN | 0.21 [0.40, 0.02] |
| Non-grammatical | ***DCF+GN > SC | 0.35 [0.18, 0.52] |
| | ***DCF+GN > DCF+G | 0.28 [0.11, 0.45] |
| | ***DCF+GN > ME+G | 0.31 [0.15, 0.48] |
| | ***ME+GN > SC | 0.43 [0.26, 0.60] |
| | ***ME+GN >DCF+G | 0.36 [0.19, 0.52] |
| | ***ME+GN >ME+G | 0.39 [0.23, 0.56] |
| Accuracy development[a] | | |
| Grammatical | ***DCF+G > SC | -0.11 [-0.16, -0.06] |
| | ***ME+G > SC | -0.06 [-0.11, -0.02] |
| | ***DCF+GN > SC | -0.09 [-0.14, -0.04] |
| | **DCF+G > ME+GN | -0.06 [-0.01, -0.11] |
| Non-grammatical | ***DCF+GN > SC | -0.09 [-0.14, -0.03] |
| | **DCF+GN > DCF+G | -0.06 [-0.12, -0.01] |
| | **DCF+GN >ME+G | -0.07 [-0.13, -0.01] |
| Perceived cognitive load | ***ME+GN > DCF+G | 1.91 [0.95, 2.88] |
| | *ME+GN > DCF+GN | 1.08 [0.09, 2.07] |
| | **ME+G > SC | 1.23 [0.24, 2.21] |
| | *DCF+GN > SC | 1.11 [0.11, 2.10] |
| | ***ME+GN > SC | 2.19 [1.21, 3.18] |
| Attitudinal engagement | | |
| Comprehensibility | *DCF+G > ME+G | 1.99 [3.81, 0.16] |
| | *DCF+G > ME+GN | 1.82 [3.63, 0.02] |
| | *DCF+GN > ME+G | 2.04 [0.20, 3.88] |
| | *DCF+GN > ME+GN | 1.88 [3.73, 0.03] |

*Note.* DCF+G = direct CF on grammatical errors; ME+G = metalinguistic CF with codes on grammatical errors; DCF+GN = direct CF on grammatical and non-grammatical errors; ME+GN = metalinguistic CF with codes on grammatical errors and non-grammatical errors; SC = self-correction with no feedback provided. * $p < .05$. ** $p < .01$. *** $p < .001$.
[a]The lower the accuracy measure obtained the more accuracy achieved.

The mixed-effect model also revealed a significant interaction effect for condition and time for overall non-grammatical accuracy, $X^2_4 = 21.24$, $p < 0.001$, $R^2 = .66$, $\Omega^2 = .62$ (see Appendix S9 and S12 in Supporting Information for full model and a summary of the significant fixed effects kept in the models, respectively). The non-grammatical accuracy of

DCF+GN was significantly better over time than that of SC ($p < 0.001$, $SE = 0.021$), DCF+G ($p = 0.008$, $SE = 0.021$), and ME+G ($p = 0.004$, $SE = 0.021$) (see Tables 2 and 3).

*5.4.Feedback effect on perceived cognitive load*

The mixed-effect model showed a statistically significant effect of condition on reported mental effort after text revision, $F(4,134) = 12$, $p < 0.001$, $R^2 = .26$, $R^2$ adjusted = .24 (see Appendix S10 and S12 in Supporting Information for full model and a summary of the significant fixed effects kept in the models, respectively). Based on self-reporting, the cognitive load imposed by revision after DCF+G was significantly lower than that after ME+GN ($p < 0.001$, $SE = 0.350$). The DCF+GN group also yielded a significantly lower cognitive load rating than the ME+GN group ($p = 0.023$, $SE = 0.356$). Similarly, the reported cognitive load of self-correcting errors with no feedback available (i.e., SC) was significantly lower than correcting them with ME+G ($p = 0.006$, $SE = 0.356$), DCF+GN ($p = 0.020$, $SE = 0.360$), and ME+GN ($p < 0.001$, $SE = 0.356$) (see Tables 2 and 3).

*5.5.Feedback effect on attitudinal engagement*

The mixed-effect model did not yield a statistically significant effect of condition on *utility* or *burden*, yet it did for *comprehensibility*, $F(3,107) = 5.17$, $p = 0.002$, $R^2 = .15$, $R^2$ adjusted = .12 (see Appendix S11 and S12 in Supporting Information for full model and a summary of the significant fixed effects kept in the models, respectively). Learners correcting errors with DCF+G reported a significantly more favorable attitude pertaining to comprehensibility than those correcting errors with ME+G ($p = 0.026$, $SE = 0.698$) and ME+GN ($p = 0.046$, $SE = 0.692$). Similarly, learners in the DCF+GN reported understanding the feedback significantly more than those in the ME+G ($p = 0.022$, $SE = 0.703$) and ME+GN ($p = 0.044$, $SE = 0.708$) groups (see Tables 2 and 3).

## 6. Discussion

This section interprets the results and touches upon the contribution and the pedagogical/theoretical implications emerging from this study.

*6.1.*To what extent did comprehensive feedback forms lead to improved accuracy during text revision and in new writings over time*?*

Results from text analyses showed a significant effect for condition during text revision as well as a significant interaction effect for condition and time in new writings (RQ1). Similar gains were also found in Truscott and Hsu (2008) and van Beuningen et al. (2008, 2012). For example, although they did not show improvement over time, the EFL learners in Truscott and Hsu's study could significantly enhance revision of the same text. Also, Dutch secondary pupils in van Beuningen et al. performed significantly better in new writings produced one (2008) and four (2012) weeks after feedback provision. A novelty in our study, though, is the evidence it provides about the different degrees of effectiveness of various comprehensive CF forms as editing (i.e., the L2 writing perspective) and learning (i.e., the SLA perspective) tools in the L2 writing class. Interpreted from a cognitive perspective (e.g., Schmidt, 1990), our findings suggest that learners were able to handle the

attentional demands of comprehensive CF: they could attend the feedback, notice (with understanding) the gap between the input (in the form of CF) and their output, match the input with their existing stored linguistic knowledge, process it, and produce accurate, modified L2 output in new writings (for stages of cognitive processing of input, see Gass, 1997 in Bitchener and Storch, 2016).

We had hypothesized that learners would be able to retrieve in new writings the knowledge that they gained from the input (i.e., written CF) and that they practiced during text revision because "we can better remember what we have learned if the cognitive processes that are active during learning are similar to those that are active during retrieval" (Lightbown, 2008, p.27). In this respect, the analysis mostly supports our hypotheses pertaining to short-term and long-term grammatical and non-grammatical accuracy gains (H1 and H2): those conditions that tapped into learners' grammatical knowledge yielded grammatical improvement (e.g., DCF+G, ME+G, DCF+GN, and ME+GN), and those that tapped into learners' non-grammatical knowledge yielded non-grammatical improvement (e.g., DCF+GN and ME+GN). Similarly, learners without attention to grammatical issues lacked grammatical improvement (e.g., SC), and learners without attention to non-grammatical issues lacked non-grammatical improvement (e.g., DCF+G, ME+G, and SC). Hence, in agreement with Schmidt (2001), our results indicate that L2 learners may not notice features they are not consciously asked to pay attention to, reducing in turn the probabilities for L2 learning to take place. Furthermore, running counter with Skehan's (1998) limited processing capacity model and in line with Robinson's (2005) positive outlook on the likely outcomes resulting from cognitively complex tasks, learners' grammatical accuracy did not suffer when their attention was also drawn to non-grammatical issues (or vice versa). The fact that simultaneous attention to multiple errors (grammatical only or both grammatical and non-grammatical) yielded evidence of short- and long-term L2 learning not only suggests that L2 learners may have enough attentional resources to cope with comprehensive corrections as evidenced in previous studies (e.g., Bonilla et al., 2017; van Beuningen et al., 2008, 2012) but also lends support to claims that the "attentional capacity problem might be more prominent in the online processing of oral feedback than in the offline handling of written CF" (van Beuningen, 2010, p.11).

Our findings suggest that differences in groups' performance could be attributed to the explicitness of the feedback type (i.e., direct CF and metalinguistic CF) and error type (i.e., grammatical and non-grammatical). Firstly, taken together, the fact that beyond text revision the metalinguistic group on grammatical and non-grammatical issues lost its advantage over the control group and that irrespective of feedback scope the direct CF groups were more effective in promoting short- and long-term L2 improvement than the metalinguistic groups, could be interpreted as further support to the claim that what matters is "the explicitness of the feedback (i.e., whether its corrective force is clear)" (Sheen, 2010, p.225). Our conclusion concurs with other researchers whose results involving direct corrections prompted them to point out the relevance of the saliency of corrective information (i.e., its explicitness) to determine the effectiveness of written CF (e.g., Nassaji & Swain, 2000; Santos, López-Serrano, & Manchón, 2010). The implied superiority of direct corrections over metalinguistic codes present in our results contributes to substantiating previous claims about direct CF

being more beneficial due to its explicitness and immediacy (cf. Chandler, 2003; Bitchener & Knoch, 2008; Ferris, 2009; Ferris et al., 2013). Secondly, similar to van Beuningen (2012), grammatical and non-grammatical errors responded differently to treatment over time. While in revised texts both grammatical and non-grammatical errors proved amenable to CF with direct corrections or metalinguistic codes, in the long term non-grammatical accuracy was durable with direct corrections only whereas grammatical issues maintained the feedback effect with either feedback type—although direct corrections proved superior. Our results coincide with those in van Beuningen et al. (2012) in that direct CF may be more advantageous to enhance grammatical accuracy in the long run, but they do not support the claim that codes may be more beneficial to remedy non-grammatical issues. Such difference in findings can be explained in light of students' confidence, which contrary to learners in our study, Beuningen et al. (2012) believed was strong enough for their students to self-correct their non-grammatical errors with codes. Also, the non-grammatical measures in the two studies may not be comparable after all because while their ratio included "lexical errors, orthographical errors, appropriateness/pragmatic errors, and other non-grammatical errors" (van Beuningen et al., 2012, p.17) , ours was computed with errors in mechanics only[5]. Clearly, our results add to previous evidence of the editing and language learning potential of comprehensive CF, yet further research is warranted on the amenability to correction of grammatical and non-grammatical errors. Furthermore, we do not discard the possibility that learner type (i.e., low-intermediate, novice writers) may have played a role in the results. Although learners in our study were low intermediate learners, they were also novice writers enrolled in a first-year course. Thus, the lack of a more advanced proficiency level combined with a lack of an advanced training in self-editing abilities as first-year students may have contributed to the extent to which learners benefited more from direct corrections than metalinguistic ones and the degree to which learners in experimental groups obtained (grammatical and non-grammatical) accuracy gains whereas those in control group did not. Ellis (2009a) mentions that without the proper linguistic knowledge, leaners are unlikely to self-correct or that sometimes learners simply prefer being corrected. It is possible then that the linguistic repertoire of the learners in our study—while still enough to benefit from codes during text revision and in new writings—may not have been sufficient to profit from codes more than from direct corrections. Further, learners' incipient knowledge of self-editing strategies (even after having reread and rewritten their text to the best of their abilities) may have proved insufficient due to their lack of training. As for learners in the control group, it also is plausible that they simply lacked the ability to detect and self-correct their errors as it is believed of learners lacking the proficiency level and/or training to do so (e.g., Polio, Fleck, & Leder, 1998).

Interestingly, a potentially effective feedback procedure for administering direct corrections may have emerged from this study. To illustrate, because in the EFL context of this investigation revision is important and passive copying of direct corrections undesirable, our feedback procedure allowed learners to study the feedback to draw their attention to form but did not let them have the feedback available while revising (also cf. Bonilla L. et al., 2017)—a procedure that may not be common L2 classroom practice. Still, operationalized in this way, our results showed that direct corrections were effective to enhance learners'

grammatical and non-grammatical accuracy in revised and new texts (even more than metalinguistic codes). As a result, it may well be that L2 teachers have a new viable alternative for correcting learners' written errors with direct corrections and in a way that may afford opportunities for language reflection, that could yield a lasting effect beyond text revision, and that may not represent a pedagogical concern. The pedagogical implication of this suggested feedback practice is noteworthy considering previous evidence concerning codes being too cryptic for L2 learners (Ferris, 1995; Hedgcock & Lefkowitz, 1996) or not being L2 learners' preferred feedback strategy when their language proficiency level is low (e.g., Bonilla et al., 2017). If L2 writing teachers do not always have the expertise (Ferris & Roberts, 2001; Truscott, 2001) or the time (Ferris, 2010) to label learners' written errors and if codes pose significantly more comprehensibility issues than direct CF—as our results indicated, L2 teachers and SLA researchers alike might want to further explore with direct corrections as operationalized in this study. Doing so would be a valuable attempt to substantiate (or not) our findings and to advance our theoretical and practical knowledge of the error correction practice.

*6.2.* What was learners' self-perceived cognitive load after revision with comprehensive CF forms*?*

Results from the mental-effort scale (Paas, 1992) rendered a statistically significant effect of condition on reported mental effort after revision with comprehensive CF forms. The analysis mostly confirmed our hypothesis (H3): the more cognitively complex the feedback processing form, the higher learners' cognitive load estimates. This was evidenced in the cline that, despite a small dent, ran as expected in the hypothesis (cf. Appendix S10).

We can explain the expected pattern of significance in light of the general model of Cognitive Load (Paas & Van Merriënboer, 1994a) and Robinson's (2001) definition of task demands. First, from the theoretical perspective of the general model of Cognitive Load, a high cognitive load is the result of complex cognitive tasks that are usually associated with a high mental load (i.e., the task-related dimension) and which in turn, tend to yield a high mental effort (i.e., the subject-related dimension). Thus, we find some explanation as to why the feedback processing forms that were associated with high mental load due the higher cognitive demands they placed on learners (e.g., metalinguistic CF with codes on grammatical and non-grammatical errors) rendered a greater perceived cognitive load than those that were associated with a lower mental load (e.g., direct CF on grammatical errors). Second, the fact that learners' cognitive load estimates tended to increase as task demands did can also be explained bearing in mind Robinson's (2001) definition of task demands. According to Robinson, the mental work load that learners engage in when performing a task increases depending on the demands the task imposes on learners. Therefore, it is likely that the mental work load (*mental load* in Paas & Van Merriënboer, 1994a) of performing what were considered more cognitively complex feedback processing forms increased because those tasks had more attentional, memory, and reasoning demands than the feedback processing forms thought to be less cognitively complex. If "tasks differ in the demands they make on our attention" (Robinson, 2003, p.642), some feedback processing forms may have

been more attention demanding than others, increasing the mental work load and rendering, in turn, higher cognitive load self-reports.

Furthermore, there may have been a cognitive difference between correcting errors with direct comprehensive CF forms and correcting them with their metalinguistic counterparts. That would explain why the former was significantly less overloading than the latter, on the one hand, and why (unlike the accuracy results) no significant differences were found between groups of the same feedback type but differing feedback scope (e.g., DCF+G and DCF+GN), on the other hand. For example, in explaining why direct corrections may be more beneficial for internalizing correct forms, Chandler (2003) posits that the cognitive expenditure of correcting one's errors may be greater. Such interpretation adds support to our results concerning the cognitive load estimates of different forms of comprehensive CF. That is, the metalinguistic CF types employed in this study entailed working out not only the meaning of the codes but also the expected target language form. Therefore, it is plausible that the problem-solving nature of the metalinguistic CF types may have placed more attentional demands on learners than the explicit provision of correct forms of the direct CF ones, which may have in turn increased their cognitive load. This interpretation is worth pursuing in future research due to its theoretical and practical repercussions: it hints at the possibility that what may be too overburdening for learners to attend to is not a broad feedback scope (as recurrently mentioned in the literature) but a low degree of feedback explicitness. Also, although our emphasis was on feedback scope due to claims about the cognitive burden of comprehensive CF, it is plausible that the cognitive strain of correcting grammatical errors may be different than that of non-grammatical errors. If error type plays a major role in its responsiveness to written CF as our study and previous research evidence have shown (e.g., Storch & Wigglesworth, 2010), it follows that different types of errors impose a different cognitive load depending on their degree of complexity and the cognitive demands that correcting them entail. Certainly, further research addressing this potential cognitive difference would be a valuable addition to the feedback literature.

6.3. What effect did comprehensive feedback forms have on learners' attitudinal engagement?

Based on the increasing complexity of the different comprehensive CF forms, we had expected some to render a more favorable attitude than others in terms of *utility*, *comprehensibility*, and *burden* (H4). However, findings from the adapted questionnaire of attitudinal engagement (cf. Bonilla L. et al., 2017) showed a statistically significant difference for comprehensibility only; learners' attitudinal engagement pertaining to utility or burden did not reach statistical significance. In fact, learners reported a similarly high score for the former and a similarly low score for the latter (cf. Table 1). Thus, the hypothesis was only partially supported.

To understand potential reasons why learners' attitudinal engagement concerning utility and burden was similar, we cannot overlook the type of learners in this study and the instructional context that surrounded the feedback: English or English Teaching majors within an FL setting. In Hedgcock and Lefkowitz´s (1994) description of L2 learners, they made a distinction between ESL learners and their FL counterparts. They claimed that FL students may not be as motivated to attend to written CF as ESL ones because the former may

view composing as product-centered, which could make them less concerned with grammatical accuracy. However, such characteristic may not be true for all FL learners. For example, an exception worth noting are those majoring in the target language. Clearly, their language learning goals, their purpose to undertake writing, and their motivation to attend to CF cannot be assumed to be the same as, for example, those from FL learners majoring in other fields and enrolling a FL course where their success in the major and future career are not necessarily at stake due to poor command of the language (e.g., Sampson, 2012; Semke, 1984). It could be argued then that as English (Teaching) majors, our participants placed a high priority on grammatical and non-grammatical accuracy, which in turn prompted them to welcome different forms of comprehensive CF more than other types of students in other contexts would. This may have enhanced any feeling of utility and lessened any feeling of emotional overload. Our results are not in line with previous descriptive studies which show that L2 learners may render a favorable emotional response despite showing signs of frustration (e.g., the FL and SL learners in the US in Hedgcock and Lefkowitz, 1994), yet they are in agreement with two feedback studies carried out in non-English dominant countries with EFL learners and which obtained evidence of learners' emotional response to written CF being favorable without signs of frustration (e.g., the distance learners in Hyland, 2001; the English majors in Bonilla et al., 2017). Interestingly, learners' reactions to comprehensive CF in particular in Sampson (2012) was not positive. Contrary to our participants, five EFL students in a Colombian university found the comprehensive treatment "discouraging" (Sampson, 2012, p.500), yet they were Economics, Finance, and Accounting majors. Thus, their motivation to engage with L2 writing and welcome CF may not have been as strong as the English (Teaching) majors in our study, further corroborating that what could make a difference in learners' reactions to written CF are learner (Hyland, 1998) and contextual variables (Hedgcock & Lefkowitz, 1996).

Another significant implication is that feedback type may be one more influential variable in how learners affectively cope with comprehensive CF. For example, a finding that was indeed consistent with the hypothesis (H4) was the significant difference between the direct CF and the ME groups pertaining to comprehensibility. Part of the theoretical basis for such expectation were the different arguments that have been advanced in favor of direct CF (cf. Bitchener, 2008; Bitchener & Knoch, 2010b; Chandler, 2003). One argument in particular states that when learners receive written CF with codes, they could have difficulties working out the corrections because they either forget the meaning of the codes or have comprehensibility issues (e.g., Ferris, 1995). Thus, this could explain why irrespective of error type, both direct groups reported understanding the feedback significantly more than either metalinguistic counterpart.

## 7. Limitations and future work

Despite the ecological validity of the present study, it was also limited by its instructional context. For example, even though this six-week feedback study did have a control group, it was not possible to maintain such condition over a longer period of time given the course demands and learners' pedagogical needs and goals. We concur with Bitchener and Ferris (2012) in that the criticism that a feedback study faces for not having a control group and, at

the same time, the ethical concern of including one when the study is "contextualized within the day-to-day activities of a writing class" (p.110), is a catch-22 situation that calls for a redefinition of *control group* (see Bitchener & Ferris, 2012, p.111 for suggestions). Unless teachers/researchers endeavor to design group contrasts that are both pedagogically feasible and methodologically acceptable, evidence of the (more) longitudinal effect of written CF within an actual writing setting will likely remain hard to obtain.

Besides, although this study was a first step into examining the cognitive load imposed by comprehensive CF, we were unable to employ an online physiological measure. Hence, to determine to what extent our findings hold under other conditions, a future research agenda may want to look into the cognitive load of feedback with a measure of eye activity, which may be feasible to implement in an L2 writing/learning environment. The advantage of eye tracking is, according to Sweller et al. (2011), that it indicates where and for how long the focus of attention is, which are indicators of variations of cognitive load. Therefore, given that longer eye fixations have proven to reflect more cognitive processing (Sweller et al., 2011), this online measure could further our (incipient) knowledge on the cognitive load of different written CF types. A case study investigation in this direction could be a good starting point.

In addition, despite the inclusion of two feedback sessions (vis-à-vis a one-shot treatment) in the design, contextual reasons prevented us from having revision on more than one occasion. Therefore, further research attempts might want to examine the durability of the feedback effect when two (or more) consecutive revision tasks are involved. Also, the present study was carried out with EFL writers within a learning-to-write and writing-to-learn language setting, so caution must be exercised when interpreting its findings.

## 8. Conclusion

The present study with a blended design sought to address unexplored accuracy-, cognitive-, and attitudinal-related issues on comprehensive CF. On the whole, it adds theoretically and practically to previous L2 writing and SLA work on written CF in a number of ways. First, pertaining to accuracy, in spite of the impossibility to counter-balance the tasks, clear differences were seen across groups in the same time and in the tendency of development. Our results suggest that for claims about comprehensive CF to be made, the feedback scope may not be the only variable to consider; other factors that could also play a role in how successfully (or not) learners cope with CF on multiple errors could be feedback explicitness, error type, and learner type. From a pedagogical standpoint, our results show that drawing leaners' attention to grammatical and non-grammatical issues simultaneously is not counterproductive when aiming for either type of accuracy. However, to further maximize grammatical accuracy, having learners attend to grammatical issues only may be a more worthwhile feedback practice. Also, while both direct corrections and metalinguistic codes seem effective to enhance short-term (grammatical and non-grammatical) accuracy, for developing learners' L2 grammatical and non-grammatical knowledge, direct corrections may have the upper hand.

Finally, our findings about the cognitive load of and learners' attitudinal engagement with comprehensive CF add a further dimension to our current understanding of learners' cognitive and affective response to such practice. We did not find any evidence of processing comprehensive CF forms being overloading (Sheen et al., 2009) or unwelcome (Truscott, 1996) so much so that corrections could not be processed and L2 learning could not take place, yet learners' cognitive load estimates proved significantly lower when corrections were provided directly and they targeted grammatical issues only. Overall, our results add theoretically and practically to the literature by suggesting that the cognitive and attitudinal response triggered by comprehensive CF is not solely determined by such feedback scope. That is, while a variable such as feedback explicitness may cause a significant difference in learners' perception of cognitive load and understanding of the feedback, learner type and instructional context could be influential factors in how learners construe how useful or emotionally burdening a given treatment could be.

**Notes**

1 Despite their unfocused group targeted a narrow number of linguistic categories, Ellis et al. (2008) and Sheen et al. (2009) construed their baseline comparison as one between focused/selective and unfocused/comprehensive CF. As a result, their studies have typically been included in discussions of comprehensive CF. However, to this day, the extent to which their treatment was comprehensive enough for such a categorization is questionable. Currently, based on Liu and Brown's (2015) classification of feedback scope, the comparison in Ellis et al. (2008) and Sheen et al. (2009) is more suited for one between highly selective and mid-selective, instead.

2 Attitudinal engagement refers to "how learners respond attitudinally to the CF" (Ellis, 2010, p.342).

3 A reviewer brought to our attention the fact that after having participated in one feedback round (i.e., in S2), learners may have been aware that the feedback would be taken away for a second round (i.e., in S4). This could have indeed posed a problem if our design had included a second revision session, but it was not the case (cf. Figure 1).

4 Cognitive load researchers have used mainly two types of indices to measure cognitive load: subjective measures (rating scales) and physiological measures (brain, heart, eye activity). The instructional context of this study prevented us from employing the latter. Still, Paas's (1992) scale has been widely adapted given its proven validity, unobtrusiveness, and easy availability (for a review of studies, see Paas et al., 2003; van Gog & Paas, 2008).

5 We also concur with a reviewer in that non-grammatical errors may have benefited more from direct corrections because stylistic errors may be less meaningful and their form may not be salient.

**References**

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). Retrieved from 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., & Ferris, D. (2012). *Written corrective feedback in second language acquisition and writing*. Routledge.

Bitchener, J., & Storch, N. (2016). *Written Corrective Feedback for L2 Development* (Vol. 96). Multilingual Matters.

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

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

Chandler, J. (2003). The efficacy of various kinds of error feedback for improvement in the accuracy and fluency of L2 student writing. *Journal of Second Language Writing*, *12*(3), 267–296. https://doi.org/10.1016/S1060-3743(03)00038-9

De Leeuw, E. D., & Dillman, D. A. (2008). *International handbook of survey methodology*. New York, NY: Taylor & Francis.

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. (2009a). Corrective feedback and teacher development. *L2 Journal*, *1*(1), 3-18.

Ellis, R. (2009b). A typology of written corrective feedback types. *ELT Journal*, *63*(2), 97–107. https://doi.org/10.1093/elt/ccn023

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, 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

Evans, N. W., Hartshorn, K. J., McCollum, R. M., & Wolfersberger, M. (2010). Contextualizing corrective feedback in second language writing pedagogy. *Language Teaching Research*, *14*(4), 445–463. https://doi.org/10.1177/1362168810375367

Ferris, D. (1995). Student Reactions to Teacher Response in Multiple-Draft Composition Classrooms. *TESOL Quarterly*, *29*(1), 33. https://doi.org/10.2307/3587804

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). New York: Cambridge University Press.

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.

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

Hartshorn, K. J., Evans, N. W., Merrill, P. F., Sudweeks, R. R., Strong-Krause, D., & Anderson, N. J. (2010). Effects of Dynamic Corrective Feedback on ESL Writing Accuracy. *TESOL Quarterly*, *44*(1), 84–109. https://doi.org/10.5054/tq.2010.213781

Hedgcock, J., & Lefkowitz, N. (1994). Feedback on feedback: Assessing learner receptivity to teacher response in L2 composing. *Journal of Second Language Writing*, *3*(2), 141–163.

Hedgcock, J., & Lefkowitz, N. (1996). Some input on input: Two analyses of student response to expert feedback in L2 writing. *The Modern Language Journal*, *80*(3), 207–308.

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. (1998). The impact of teacher written feedback on individual writers. *Journal of Second Language Writing*, *7*(3), 255–286. https://doi.org/10.1016/S1060-3743(98)90017-0

Hyland, F. (2001). Providing effective support: Investigating feedback to distance language learners. *Open Learning*, *16*(3), 233–247. https://doi.org/10.1080/02680510120084959

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

Hyland, K., & Hyland, F. (2006). Feedback on second language students' writing. *Language Teaching*, *39*(02), 83. https://doi.org/10.1017/S0261444806003399

Incecay, V., & Dollar, Y. K. (2011). Foreign language learners' beliefs about grammar instruction and error correction. *Procedia - Social and Behavioral Sciences*, *15*, 3394–3398. https://doi.org/10.1016/j.sbspro.2011.04.307

Kang, E., & Han, Z. (2015). The Efficacy of Written Corrective Feedback in Improving L2 Written Accuracy: A Meta-Analysis. *The Modern Language Journal*, *99*(1), 1–18. https://doi.org/10.1111/modl.12189

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

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

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

Leppink, J., Paas, F., Van der Vleuten, C. P. M., Van Gog, T., & van Merriënboer, J. J. G. (2013). Development of an instrument for measuring different types of cognitive load. *Behavior Research Methods*, *45*(4), 1058–1072. https://doi.org/10.3758/s13428-013-0334-1

Lightbown, P. M. (2008). Transfer appropriate processing as a model for classroom second language acquisition. In H. Zhaohong (Ed.), *Understanding second language process* (pp. 27–44). Clevedon, UK: Multilingual Matters.

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

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

Manchón, R. M. (2011). *Learning-to-write and writing-to-Learn in an additional language*. Amsterdam: John Benjamins.

Murphy, L., & Roca de Larios, J. (2010). Feedback in second language writing: An introduction. *International Journal of English Studies*, *10*(2), 0–I.

Nassaji, H., & Swain, M. (2000). A Vygotskian Perspective on Corrective Feedback in L2: The Effect of Random Versus Negotiated Help on the Learning of English Articles. *Language Awareness*, *9*(1), 34–51. https://doi.org/10.1080/09658410008667135

Ortega, L. (2012). Epilogue: Exploring L2 writing–SLA interfaces. *Journal of Second Language Writing*, *21*(4), 404–415. https://doi.org/10.1016/j.jslw.2012.09.002

Paas, F. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive-load approach. *Journal of Educational Psychology*, *84*(4), 429–434.

Paas, F., Tuovinen, J. E., Tabbers, H., & Van Gerven, P. W. M. (2003). Cognitive Load Measurement as a Means to Advance Cognitive Load Theory. *Educational Psychologist*, *38*(1), 63–71. https://doi.org/10.1207/S15326985EP3801_8

Paas, F., & van Merriënboer, J. J. G. (1994). Instructional control of cognitive load in the training of complex cognitive tasks. *Educational Psychology Review*, *6*, 51–71.

Paas, F., & van Merriënboer, J. J. G. (1994). Variability of worked examples and transfer of geometrical problem-solving skills: A cognitive-load approach. *Journal of Educational Psychology*, *86*(1), 122.

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

Polio, C., Fleck, C., & Leder, N. (1998). "If I only had more time:" ESL learners' changes in linguistic accuracy on essay revisions. *Journal of Second Language Writing*, *7*(1), 43–68.

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

Robinson, P. (1995). Attention, memory and the "noticing" hypothesis. *Language Learning*, *45*, 283–331.

Robinson, P. (2001). Task complexity, cognitive resources, and syllabus design: A triadic framework for examining task influences on SLA. In P. Robinson (Ed.), *Cognition and Second Language Instruction* (pp. 285–318). Cambridge: Cambridge University Press.

Robinson, P. (2003). Attention and memory during SLA. In C. J. Doughty & M. H. Long (Eds.), *Handbook of second language acquisition* (pp. 631–678). Malden, Mass: Blackwell.

Robinson, P. (2005). Cognitive complexity and task sequencing: Studies in a componential framework for second language task design. *IRAL*, *43*, 1–32.

Robinson, P., & Gilabert, R. (2007). Task complexity, the Cognition Hypothesis and second language learning and performance. *IRAL - International Review of Applied Linguistics in Language Teaching*, *45*(3). https://doi.org/10.1515/iral.2007.007

Sampson, A. (2012). "Coded and uncoded error feedback: Effects on error frequencies in adult Colombian EFL learners'' writing"." *System*, *40*(4), 494–504. https://doi.org/10.1016/j.system.2012.10.001

Santos, M., López-Serrano, S., & Manchón, R. M. (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.

Schmidt, R. W. (1990). The role of consciousness in second language learning. *Applied Linguistics*, *11*(2), 129–158.

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

Semke, H. D. (1984). Effects of the red pen. *Foreign Language Annals*, *17*(3), 195–202.

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

Sheppard, K. (1992). Two Feedback Types: Do They Make A Difference? *RELC Journal*, *23*(1), 103–110. https://doi.org/10.1177/003368829202300107

Shintani, & 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

Skehan, P. (1998). *A Cognitive Approach to Language Learning*. Oxford: Oxford University Press.

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

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, *12*(2), 257–285.

Sweller, J., Ayres, P., & Kalyuga, S. (2011). *Cognitive Load Theory*. New York, NY: Springer New York.

Taber, K. S. (2017). The Use of Cronbach's Alpha When Developing and Reporting Research Instruments in Science Education. *Research in Science Education*. 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.

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

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.

van Beuningen, C., De Jong, N. H., & Kuiken, F. (2008). The effect of direct and indirect corrective feedback on L2 learners' written accuracy. *ITL-International Journal of Applied Linguistics*, *156*, 279–296.

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 Gog, T., & Paas, F. (2008). Instructional Efficiency: Revisiting the Original Construct in Educational Research. *Educational Psychologist*, *43*(1), 16–26. https://doi.org/10.1080/00461520701756248

**Supporting Information**

**Appendix S1.** Sample Coding System
**Appendix S2.** Instructions for Studying the Feedback
**Appendix S3.** Instructions for Revising the Text
**Appendix S4.** Sample Writing Task
**Appendix S5.** Mental-effort Scale
**Appendix S6.** Scale Items and Summary of Factor Loadings
**Appendix S7.** Descriptive Statistics for Preliminary Analyses
**Appendix S8.** Mixed-effect Model and Figure for Immediate Grammatical and Non-grammatical Accuracy
**Appendix S9.** Mixed-effect Model and Figure for Grammatical and Non-grammatical Development

**Appendix S10.** Mixed-effect Model and Figure for Perceived Cognitive Load
**Appendix S11.** Mixed-effect Model and Figure for Attitudinal Engagement
**Appendix S12.** Summary of Significant Fixed Effects Kept in the Models

| Error type code | Spelled out form | Brief description |
|---|---|---|
| SV | Subject-verb agreement | Subject and verb lack agreement in number |
| ART | Article | Unnecessary insertion, faulty, or missing definite or indefinite article |
| VB | Verb | Wrong formation of verb phrase or erroneous choice of tense |
| PR | Pronoun | Incorrect or missing pronoun |
| MOD | Modal | Incorrect or missing modal |
| PREP | Preposition | Faulty or missing preposition |
| WF | Word form | Faulty or missing word endings |
| SD | Subject deletion | Omission of subject in the sentence |
| SR | Subject repetition | Insertion of an unnecessary subject |
| SS | Sentence structure | Word order or unnecessary words or phrases |
| FRAG | Sentence fragment | Incomplete thoughts: omission of words, phrases, or clauses |
| SP | Spelling | Misspelled word |
| PUNCT | Punctuation | Incorrect or missing punctuation mark |
| CAP | Capitalization | Wrong or missing capitalization |

**INSTRUCTIONS:** Study carefully the copy of the text you wrote two days ago and see in which way(s) it can be improved.

**INSTRUCTIONS:**
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.

**Writing task 1**

Instructions.
Answer the following question in a 175-word opinion paragraph.

Do you agree that people should exaggerate the truth or outright lie in their resume if that will help them to get a job? Why or why not?

Explain your reasons clearly. Use examples from your own experience to support your general ideas.

Instructions: Circle the number that best fits your intensity of mental effort.

In the composition that I just finished, I invested…

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|
| very, very low mental effort | very low mental effort | low mental effort | rather low mental effort | neither low, nor high mental effort | rather high mental effort | high mental effort | very high mental effort | very, very high mental effort |

Table

*Rotated Factor Loadings for Learners' Attitudes Towards the Feedback*

| | Factor | | |
|---|---|---|---|
| Item | 1 | 2 | 3 |
| I. Utility | | | |
| 1. Were the corrections useful? | .764 | | |
| 7. Did you find the corrections ineffective? | .715 | | |
| 4.Were you able to correct your errors using the feedback? | .623 | | |
| 5. Did you feel motivated to revise the text? | .586 | | |
| II. Comprehensibility | | | |
| 9. Were the corrections legible? | | .885 | |
| 3. Were the corrections clear? | | .705 | |
| 10. Were the corrections easy to follow? | | .594 | |
| 6. Did the corrections confuse you? | | .536 | |
| III. Burden | | | |
| 2. Did the corrections frustrate you? | | | .841 |
| 8. Did the corrections overwhelm you? | | | .810 |

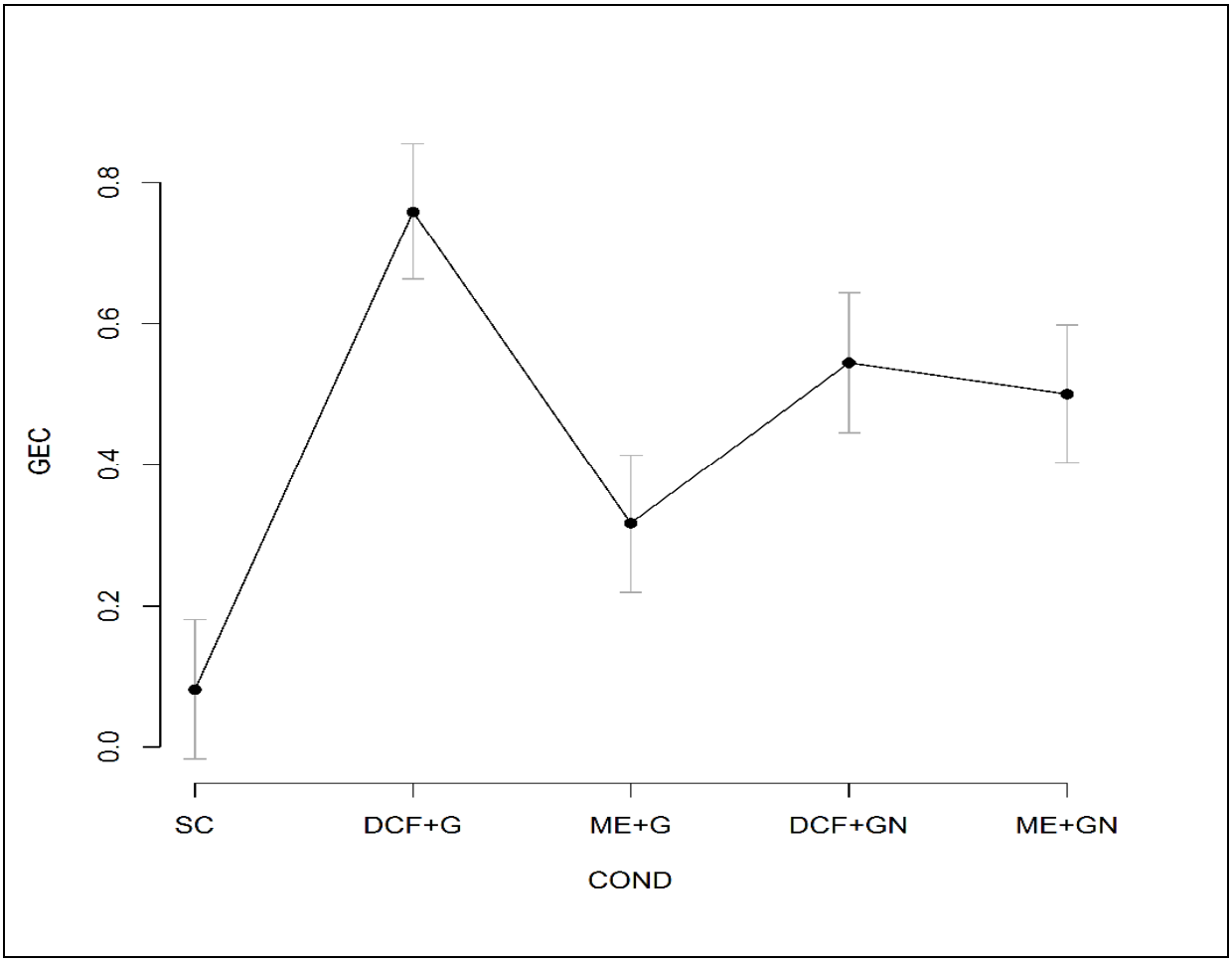
Table

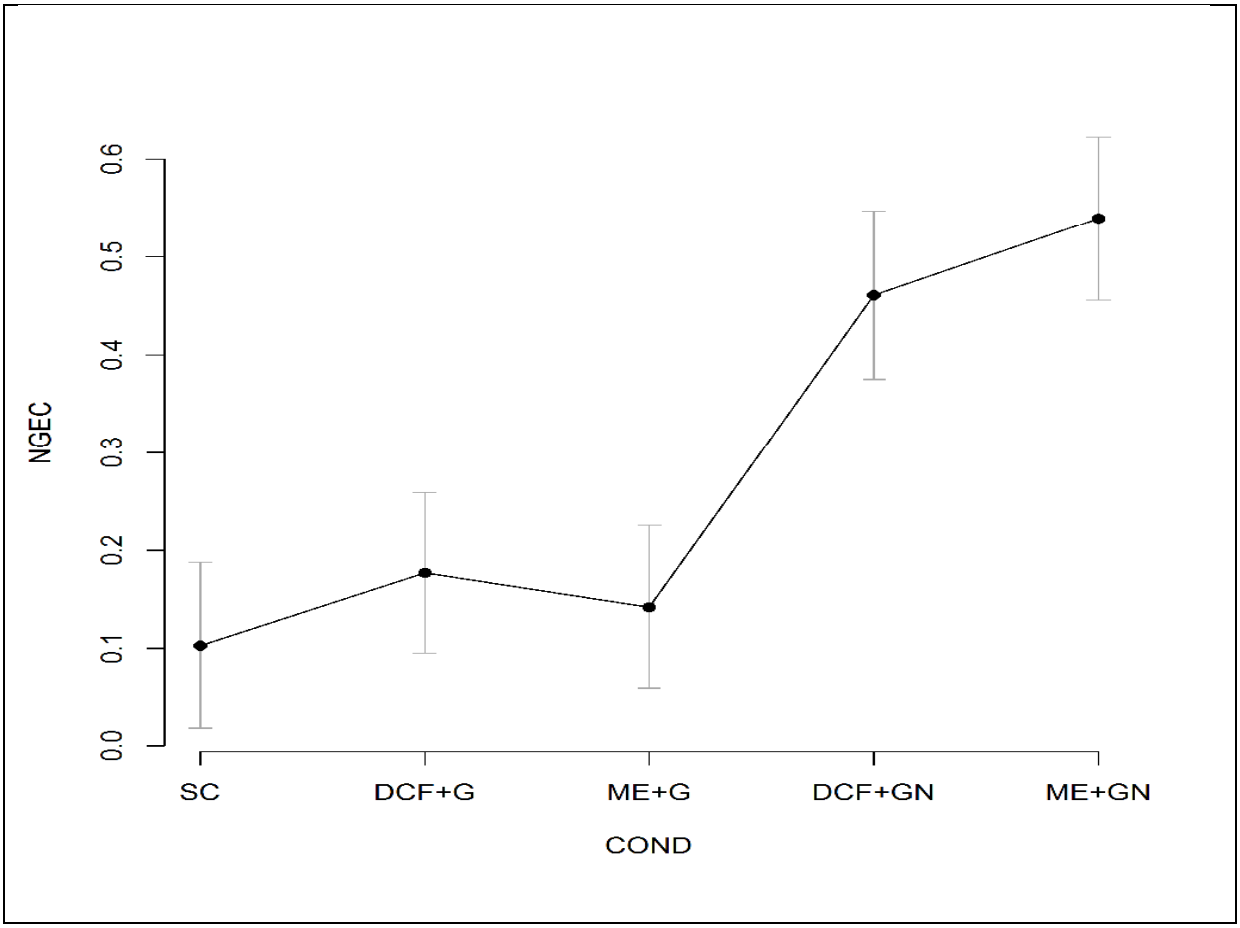
*Summary of Descriptive Statistics for Nonsignificant Differences in Preliminary Analyses*

| | | Proficiency level | | Overall grammatical accuracy | | Overall non-grammatical accuracy | | Perceived cognitive load | |
|---|---|---|---|---|---|---|---|---|---|
| Condition | *N* | *M* | *SD* | *M* | *SD* | *M* | *SD* | *M* | *SD* |
| DCF+G | 29 | 2.00 | .535 | .387 | .247 | .365 | .204 | 5.28 | 1.33 |
| ME+G | 28 | 2.36 | .826 | .325 | .237 | .309 | .167 | 5.61 | 1.42 |
| DCF+GN | 27 | 2.56 | .847 | .331 | .252 | .400 | .213 | 5.44 | 1.12 |
| ME+GN | 28 | 2.21 | .787 | .353 | .260 | .363 | .228 | 4.71 | 1.04 |
| SC | 27 | 2.26 | .903 | .312 | .261 | .349 | .207 | 5.33 | 1.20 |
| Total | 139 | 2.27 | .797 | .342 | .249 | .357 | .204 | 5.27 | 1.25 |

*Note.* DCF+G = direct CF on grammatical errors; ME+G = metalinguistic CF with codes on grammatical errors; DCF+GN = direct CF on grammatical and non-grammatical errors; ME+GN = metalinguistic CF with codes on grammatical errors and non-grammatical errors; SC = self-correction with no feedback provided.

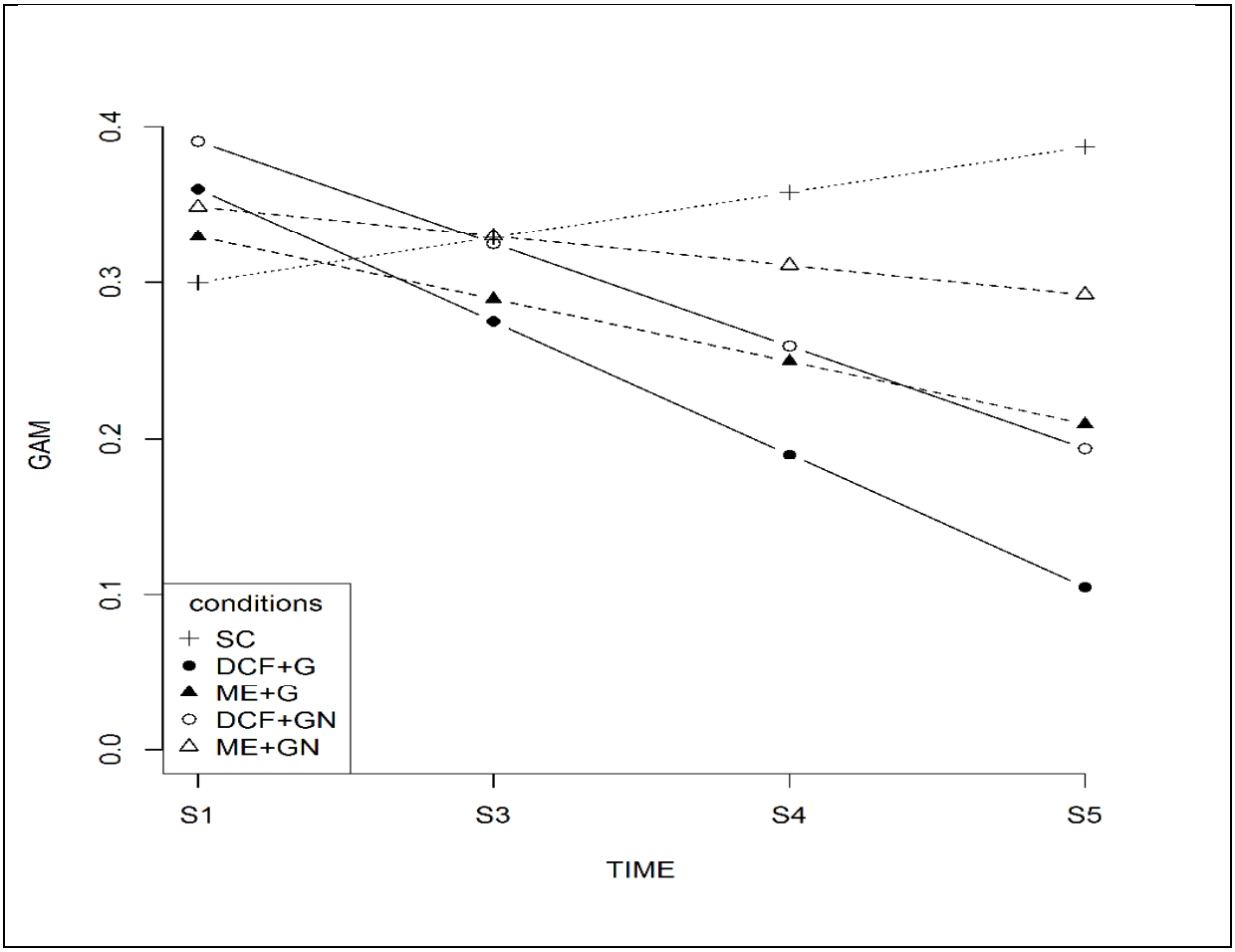
To analyze the feedback effect during text revision (RQ1), we selected the model GEC ~ TGWT1 + COND for immediate grammatical accuracy and NGEC ~ TNGWT1 + COND for immediate non-grammatical accuracy. In such models, GEC (i.e., the proportion of grammatical errors successfully corrected) and NGEC (i.e., the proportion of non-grammatical errors successfully corrected) were response variables whereas TGWT1 (i.e., the number of grammatical errors in writing task 1), TNGWT1 (i.e., the number of non-grammatical errors in writing task 1) and COND (i.e., condition) were predictors. A number of additional candidate predictors were considered for inclusion one by one but did not significantly improve the model. Hence, they were dropped. This is the case for the interaction COND:TGWT1 as well as for the variables MER (i.e., mental effort during revision) and PROFI (i.e., proficiency)—which were tested for inclusion both with and without their two-way interactions with the other predictors). The contribution of COND to the model EC ~ COND + TGWT1 is significant [$F(4,133) = 26.47$, $p < 0.0001$] as is the contribution of TGWT1 [$F(1,133) = 4.88$, $p = 0.029$]. As for the model NGEC ~ COND + TNGWT1, the contribution of COND [$F(4,133) = 22.07$, $p < 0.0001$] and TNGWT1 [($F(1,133) = 3.61$, $p = 0.059$)] was also significant. The figures below graphically depict the effect of condition in the model for successful correction of grammatical (GEC) and non-grammatical errors (NGEC), respectively.

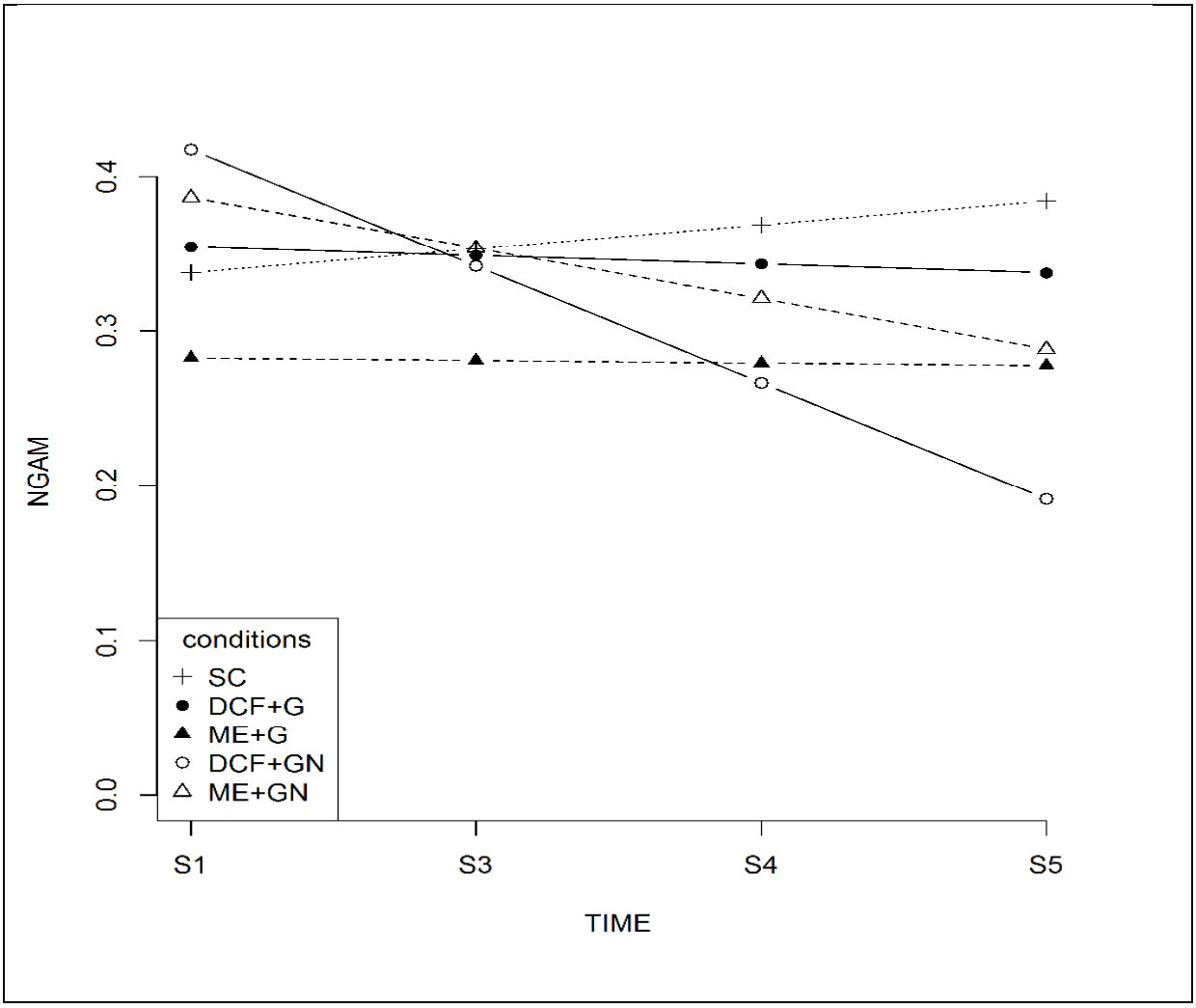



NGEC
0.0
0.1
0.2
0.3
0.4
0.5
0.6
SC
DCF+G
ME+G
DCF+GN
ME+GN
COND

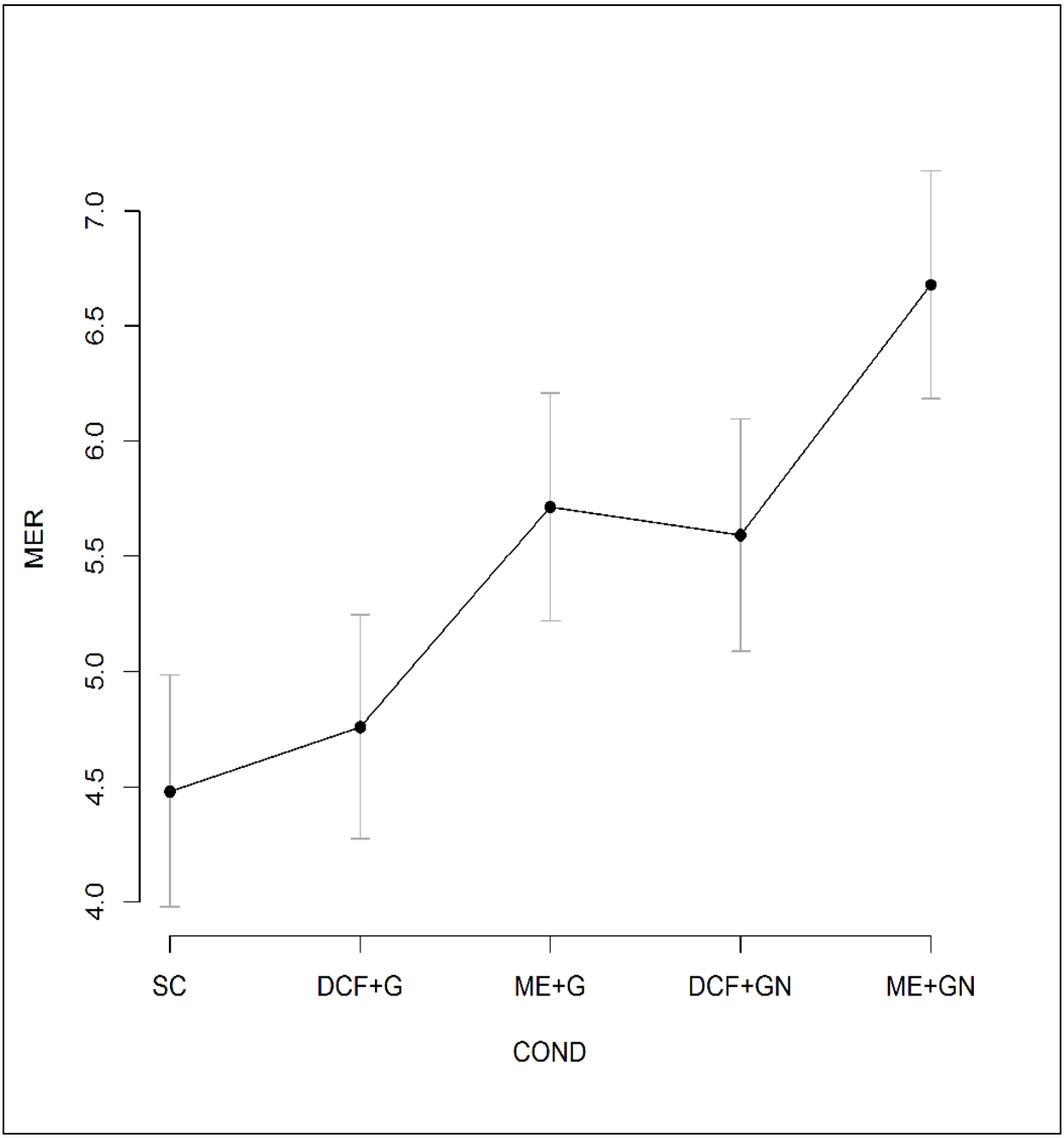
In our analysis of the effect of feedback on long-term grammatical and non-grammatical accuracy (RQ1), our data contain four observations for each participant (one for each writing task). In each observation, the variables GAM and NGAM, respectively, capture the grammatical and non-grammatical accuracy in the writing tasks; the variable TIME (treated numerically) identifies the new writing tasks (i.e., in sessions 1, 3, 4, and 5); the variable CEFR_s identifies learners' standardized CEFR scores; and the variable PARTIC.ID (with a unique ID for each participant) identifies the participant. The variable COND turned out to interact significantly with TIME. Another candidate predictor, MER, did not contribute significantly to the model, so it was not added. Thus, the resulting model for examining grammatical development was GAM ~ COND + TIME + CEFR_s + CEFR_s:TIME + COND:TIME + (1+TIME|PARTIC.ID) and for non-grammatical development, it was NGAM ~ COND + TIME + CEFR_s + CEFR_s:TIME + COND:TIME + (1+TIME|PARTIC.ID). As the random component in the models formula indicates, we added by-participant random intercepts as well as a by-participant random slope for time. For grammatical development, the random intercept has a standard deviation of 0.149; the random slope has a standard deviation of 0.020; and the residuals have a standard deviation of 0.140. For non-grammatical development, the random intercept has a standard deviation of 0.124; the random slope has a standard deviation of 0.042; and the residuals have a standard deviation of 0.147. All fixed-effect predictors in the mixed-effects model GAM ~ COND + TIME + CEFR_s + CEFR_s:TIME + COND:TIME + (1+TIME|PARTIC.ID) contribute significantly to the model. This applies to the interaction COND:TIME ($X^2_4$ = 44.31, $p$ = < 0.0001) and the interaction CEFR_s:TIME ($X^2_1$ = 19.94, $p$ = < 0.0001). The same is true for the interaction COND:TIME ($X^2_4$ = 21.24, $p$ = < 0.0001) and the interaction CEFR_s:TIME ($X^2_1$ = 9.12, $p$ = < 0.0001) in the mixed-effects model NGAM ~ COND + TIME + CEFR_s + CEFR_s:TIME + COND:TIME + (1+TIME|PARTIC.ID). The interaction effect of condition and time in the model for grammatical (GAM) and non-grammatical (NGAM) improvement, respectively, is shown in the figures below[1]; confidence limits were omitted in order not to clutter the plot.

[1] The lower the score obtained the more accuracy achieved.

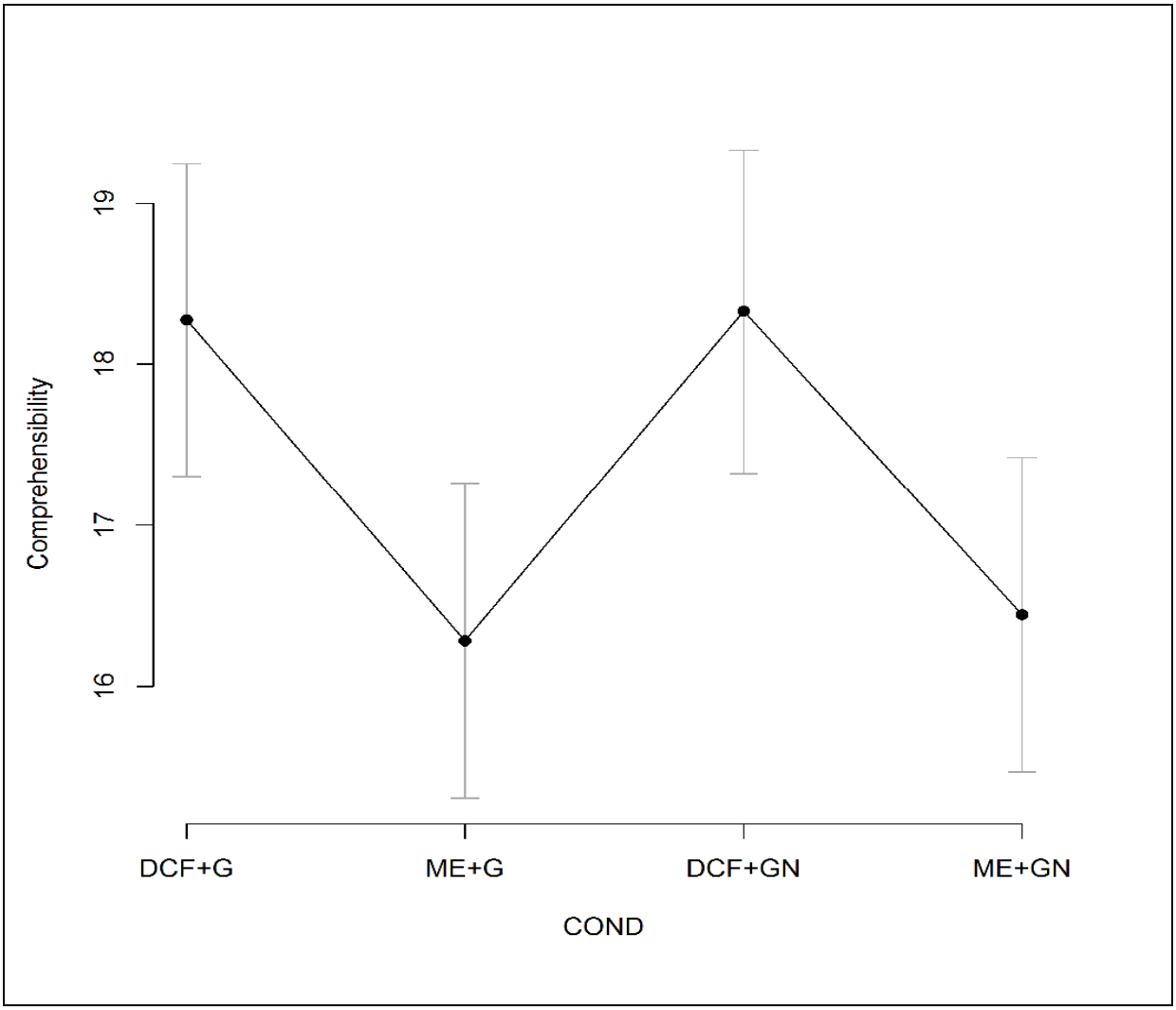

GAM
0.0
0.1
0.2
0.3
0.4
S1
S3
S4
S5
TIME
conditions
SC
DCF+G
ME+G
DCF+GN
ME+GN


NGAM
0.0
0.1
0.2
0.3
0.4
S1
S3
S4
S5
TIME
conditions
SC
DCF+G
ME+G
DCF+GN
ME+GN

For examining the effect of CF on cognitive load (RQ2), we ran a linear regression analysis MER ~ COND, with MER as response variable and COND as predictor. Neither the candidate predictor PROFI nor the interaction COND:PROFI turned out to significantly improve the model; both were kept out of the model. The contribution of COND to the model MER ~ COND is significant [$F(4,134) = 12$, $p < 0.0001$]. The figure below shows the effect of condition in the model.



For examining the feedback effect on learners' attitudinal engagement (RQ3), the variables COND, PROFI, and their interaction COND:PROFI, were tested as candidate predictors in models with Utility, Comprehensibility, and Burden as response variable. None of the candidate predictors turned out to have a significant effect on the response variables Utility or Burden. For the response variable Comprehensibility, the model that was selected, was Comprehensibility ~ CEFR_s + COND. While the contribution of CEFR_s [$F(1,107) = 5.02$, $p = 0.027$] and COND [$F(3,107) = 5.17$, $p = 0.002$] was significant, the candidate predictors PROFI and COND:PROFI did not contribute significantly and were not added. The figure below graphically depicts the effect of condition in the model[1].




[1] Learners in the SC condition were not considered in the analyses.

**Table**

*Summary of Significant Fixed Effects Kept in the Models*

| | Immediate grammatical accuracy | Immediate non-grammatical accuracy | Grammatical development | Non-grammatical development | Perceived cognitive load | Attitudinal engagement | |
|---|---|---|---|---|---|---|---|
| | | | | | | Comprehensibility | |
| | *b* [95% CI] | *b* [95% CI] | *b* [95% CI] | *b* [95% CI] | *b* [95% CI] | | *b* [95% CI] |
| Condition | | | | | | Condition | |
| SC → DCF+G | 0.67*** [0.53, 0.81] | 0.07 [-0.04, 0.19] | -0.11* [-0.19, -0.02] | -0.01 [-0.09, 0.05] | 0.27 [-0.42, 0.97] | DCF+G → ME+G | -1.99** [-3.37, -0.60] |
| SC → ME+G | 0.23** [0.09, 0.37] | 0.03 [-0.07, 0.15] | -0.07* [-0.16, 0.01] | -0.08* [-0.15, -0.00] | 1.23*** [0.52, 1.93] | DCF+G → DCF+GN | 0.05 [-1.35, 1.46] |
| SC → DCF+GN | 0.46*** [0.32, 0.60] | 0.35*** [0.23, 0.47] | -0.05 [-0.13, 0.03] | -0.05 [-0.13, 0.01] | 1.11** [0.39, 1.82] | DCF+G → ME+GN | -1.82** [-3.202] |
| SC → ME+GN | 0.41*** [0.27, 0.55] | 0.43*** [0.31, 0.55] | -0.02 [-0.10, 0.06] | -0.02 [-0.09, 0.05] | 2.19*** [1.49, 2.90] | | |
| Condition:Time | | | | | | | |
| SC → DCF+G | | | -0.11*** [-0.14, -0.07] | -0.02 [-0.06, 0.01] | | | |
| SC → ME+G | | | -0.06*** [-0.10, -0.03] | -0.01 [-0.05, 0.02] | | | |
| SC → DCF+GN | | | -0.09*** -0.12, -0.05] | -0.09*** [-0.13, -0.04] | | | |
| SC → ME+GN | | | -0.04** [-0.08, -0.01] | -0.04* [-0.08, -0.00] | | | |
| Time | | | 0.02* [0.00, 0.05] | 0.01 [-0.01, 0.04] | | | |
| CEFR_s | | | -0.08*** [-0.11, -0.05] | -0.03** [0.06, -0.01] | | | 0.57* [0.06, 1.07] |
| Time:CEFR_s | | | 0.02*** [0.01, 0.03] | | | | |

**Table** Continued

| | Immediate grammatical accuracy | Immediate non-grammatical accuracy | Grammatical development | Non-grammatical development | Perceived cognitive load | Attitudinal engagement |
|---|---|---|---|---|---|---|
| | | | | | | Comprehensibility |
| | *b* [95% CI] | *b* [95% CI] | *b* [95% CI] | *b* [95% CI] | *b* [95% CI] | *b* [95% CI] |
| TGWT1 | -0.01* [-0.02, -0.00] | | | | | |
| TNGWT1 | | 0.00 [-0.00, 0.01] | | | | |
| Intercept | 0.15** [0.03, 0.26] | 0.04 [-0.06, 0.14] | 0.34*** [0.28, 0.40] | 0.36*** [0.30, 0.41] | 4.48*** [3.97, 4.98] | 18.27*** [17.30,19.24] |
| Observations | 139 | 139 | 556 | 556 | 139 | 112 |
| Effect size | $R^2$ 0.42 | $R^2$ 0.40 | $\Omega^2$ 0.71 | $\Omega^2$ 0.62 | $R^2$ 0.24 | $R^2$ 0.12 |

*Note*. DCF+G = direct CF on grammatical errors; ME+G = metalinguistic CF with codes on grammatical errors; DCF+GN = direct CF on grammatical and non-grammatical errors; ME+GN = metalinguistic CF with codes on grammatical errors and non-grammatical errors; SC = self-correction with no feedback provided; CEFR_s = standardized CEFR scores; TGWT1 = the number of grammatical errors in writing task 1; TNGWT1 = the number of non-grammatical errors in writing task 1. * $p < .05$, ** $p < .01$, *** $p < .001$.