Received: 29 January 2021 | Accepted: 3 February 2022

DOI: 10.1111/flan.12611

ORIGINAL ARTICLE

ACTFL Language Connects WILEY

Effects of collaborative writing and peer feedback on Spanish as a foreign language writing performance

Mª Camino Bueno-Alastuey[1] | Raychel Vasseur[2] | Idoia Elola[2]

The Challenge

Collaborative writing and peer feedback have been analyzed as positive writing practices mainly in English as an L2. But are they equally effective? Are they also as positive in Spanish as a foreign language? And is there any specific order to implement both treatments which is more beneficial?

[1]Department of Humanities and Education Sciencies, I-Communitas: Institute for Advanced Social Research, Public University of Navarre (UPNA), Pamplona, Spain

[2]Department of Classical and Modern Languages and Literatures, Texas Tech University, Lubbock, Texas, USA

Correspondence
Mª Camino Bueno-Alastuey, I-Communitas: Institute for Advanced Social Research, Public University of Navarre, Pamplona, Spain.
Email: camino.bueno@unavarra.es

Abstract

This study explores the effect of collaborative writing (CW) and peer feedback (PF) practices on subsequent individual writing assignments. Two groups of university students in a Spanish as a foreign language course experienced both CW and PF (Group 1 CW then PF; Group 2 PF then CW), and pre and posttests were analyzed for syntactic complexity, lexical diversity, accuracy, and fluency, as well as for overall quality using an analytic scale. Results suggest both treatments produced improvements, although PF was more beneficial for syntactic complexity, fluency, and overall quality, while CW led to more accurate texts. The order of treatments also affected scores: PF followed by CW produced better results in overall quality and fluency, while CW followed by PF was more beneficial for syntactic complexity and accuracy. Based on the results,

pedagogical implications and recommendations, as well as limitations and suggestions for future research, are provided.

KEYWORDS
analytic scale, CAF measures, collaborative writing, holistic assessment, peer feedback, writing in Spanish as a foreign language

1 | INTRODUCTION

The concept of writing as a social act has increasingly been adopted and/or revisited, due in part to the advent of new technologies and social tools in our daily lives, and to a renewed interest in scaffolding practices (grounded in sociocultural theory [Vygotsky, 1986]), and in promoting interaction with a focus on form (grounded in interactionist perspectives [Long, 1996; Swain, 2000]) in language courses. Writing in Spanish as a foreign (FL) or heritage language[1] has increasingly integrated collaborative writing (CW)[2] and peer-feedback (PF)[3] as ways to make learning more interactive and student-centered (Elola, 2017; Oskoz & Elola, 2020). Both practices allow students to engage with problem-solving and decision-making activities to express their ideas in their second language (L2), helping them recognize their limitations or gaps when trying to match their linguistic knowledge to the demands of writing (Swain, 2000). Additionally, they allow students to reconstruct knowledge, promoting language learning (Swain, 2010), and provide a space to face linguistic and composing challenges (Kuiken & Vedder, 2002) connected to pedagogical approaches in which students have more learning agency.

Another important reason for the use of CW and PF is the opportunities they offer to create collaborative dialogs that represent language in progress (Mendonça & Johnson, 1994; Watanabe & Swain, 2007). This representation takes the form of language-related episodes (LREs) that help draw students' attention toward language form and lead to the co-construction of linguistic knowledge (Suzuki & Storch, 2020; Villarreal & Gil–Sarratea, 2020). Research has focused on analyzing the impact CW and PF practices have on final products, comparing them to individual texts, however, few studies have investigated the effects those treatments have on subsequent assignments completed individually (with exceptions such as Bikowski & Vithanage, 2016; Bueno-Alastuey & Rodero Albaiceta, 2019; Chen, 2019; Olovson, 2018; Tian, 2011). Generally, the written texts produced by students working in the two conditions have been analyzed by focusing on their complexity, accuracy, and fluency (CAF). By using these constructs, studies have attempted to conceptualize the evolving complexity of syntactic dimensions of language to inform L2 acquisition theory (Casals & Lee, 2019). However, these constructs have not always been measured using the same parameters, thus making comparisons between studies difficult and, in some ways, unsustainable or inconsistent (Polio & Shea, 2014). For example, some studies have used T-units (which consist of a main clause and all subordinate clauses and nonclausal structures attached to or embedded in it [Hunt, 1970]) to analyze syntactic complexity, while other studies have used clauses, thus, making their results difficult to compare. Inconsistencies and variety of measures have also been noted by Zhang and Plonsky (2020) in their meta-analysis of CW in face-to-face settings as

reasons for the difficulties to compare studies. The authors also reported the main focus (82% of studies) on English as an L2 or FL in intermediate and advanced level courses, indicating a lack of research in other L2s. The same issues affect PF research, especially the lack of research examining languages other than English as an L2 or FL.

One reason for the plethora of work with CAF analysis of texts written in English could be the existence of computational systems for automatic analysis, which facilitate the complex analysis of many indices (Lu, 2010). The lack of such software packages for other languages to perform complex, automatic analyses might have contributed to the scarce research and the limited amount of features and number of texts analyzed when comparing CW, PF, and individual writing assignments for other L2s. Analyzing Spanish texts and including more indices in those analyses will further our understanding of students' development in some features unique to Spanish, such as aspects of the past tense—imperfect versus preterit—that are not encountered in English.

Therefore, the aim of the current study is three-fold: (1) to see if CW and PF as treatments have any influence on subsequent, individually written Spanish assignments; (2) to see if either treatment (CW or PF) is more beneficial for subsequent individual writing assignments; and (3) to analyze texts written after those treatments using holistic ratings as well as more robust and varied CAF measures than those reported in previous studies.

2 | LITERATURE REVIEW

2.1 | Collaborative writing

CW—in which two or more people work together to produce a document with group responsibility for the end product (Bosley, 1989)—has been increasingly adopted in FL education. Through the act of writing collaboratively, learners tend to engage in a dialog that impels them to notice gaps in their L2 production and to test new hypotheses regarding language and literacy acquisition. Consequently, it expands the notion of learning from just the product of one individual's efforts to a social act that is connected to the surroundings, tools, and the overall context in which the learning takes place. Research on CW has investigated the quantity and quality of texts (Arnold et al., 2009; Kessler, 2009; Lee, 2010; Olovson, 2018; Oskoz & Elola, 2011), their accuracy, complexity, and fluency (Arnold et al., 2009; Kessler, 2009; Olovson, 2018; Oskoz & Elola, 2014), the concept of authorship (Lee, 2010), writers' composing processes (Arnold et al., 2009; Oskoz & Elola, 2014), learners' greater focus on meaning than form (Aydin & Yildiz, 2014; Kessler, 2009), and learners' reorientation and development of a shared perspective regarding the task at hand and the tools being used (Oskoz & Elola, 2012, 2013, 2014).

Additionally, studies based on student perspectives have shown positive attitudes when working in pairs and groups (Arnold et al., 2009; Storch, 2005; Villarreal & Munarriz-Ibarrola, 2021); a growth in self-confidence and improvement of their speaking abilities (Shehadeh, 2011); a sense of development in their language including, for example, their understanding of the grammatical, and lexical aspects of language (Fernández Dobao & Blum, 2013; Villarreal & Munarriz-Ibarrola, 2021); an increase in cultural awareness (Al-Jamhoor, 2005); increased awareness of their audience and the different genres (Alyousef & Picard, 2011); and students' perception of online CW both as a scaffolded and interactively guided space in which they co-created (Limbu & Markauskaite, 2015).

Other studies have concentrated on comparing CW versus individual writing, many of them measuring improvements through a CAF analysis (see, for instance, García Mayo, 2007; Villarreal & Gil–Sarratea, 2020). In a meta-analysis, Elabdali (2021) showed that CW produces more accurate texts than individual writing but the results regarding other dependent variables were not conclusive. For example, Olovson (2018) suggested CW resulted in better quality texts as measured by accuracy, complexity, and lexical quality than individual writing for L2 learners of Spanish. Storch (2005) observed higher complexity in CW texts than in individual texts, while other authors (Fernández Dobao, 2012; Strobl, 2014; Wigglesworth & Storch, 2009) claimed that working collaboratively or individually had no impact on the complexity of their texts. It is important to note that Storch (2005) and Wigglesworth and Storch (2009) measured complexity as the number of clauses to T-units and percentage of dependent clauses to a total number of clauses. Fernández Dobao (2012) similarly measured complexity using words per clause, words per T-unit, and clauses per T-unit. Strobl (2014), however, used an algorithm combining four textual surface complexity measures into one numerical value for balanced complexity. More recently, others (Bueno-Alastuey & Martínez de Lizarrondo, 2017; McDonough et al., 2018) have noted a higher proportion of subordinate clauses in texts written individually than collaboratively. When considering accuracy, Fernández Dobao (2012) showed texts written by groups were more accurate than those written by pairs, but similar in complexity and fluency, whereas texts written by groups and pairs were more accurate and shorter than those written individually, but similar in complexity. However, Bueno-Alastuey and Martínez de Lizarrondo (2017) found that texts were generally more accurate, more complex, and longer when written in groups than in pairs. Biria and Jafari's (2013) study showed that students' final essays revealed that, on average, pairs produced statistically more T-units and clauses than individuals. They also found practicing in pairs improved the overall quality of the learners' written productions. Strobl's (2014) study also found that collaboration led to higher text accuracy and more appropriate content selection and organization. When examining fluency, no significant differences have been found between collaborative and individual work (Biria & Jafari, 2013; Fernández Dobao, 2012; Storch, 2005; Strobl, 2014; Wigglesworth & Storch, 2009).

Research on the impact of CW tasks on subsequent individual assignments is scarce with the exception of Bikowski and Vithanage's (2016) and Chen's (2019) studies. The former reported statistically significant writing gains for the CW group comparing individual writing pretest and posttest scores based on an analytic rubric focused on content, organization, academic style, and grammar. Their results were corroborated by Chen (2019), whose experimental group used CW for 16 weeks, and experienced statistically significant improvements in accuracy and fluency, but not in complexity, and in quality ratings of vocabulary and grammar in subsequent individual writings using a pretest, posttest, and delayed posttest design.

2.2 | Peer-feedback

PF (also referred to as *peer response*, *peer assessment*, *peer review*, or *peer editing*) provides students the opportunity to interact by commenting orally or in written form on each other's writing, thus, opening the opportunity for discussion, communication, and collaboration in a skill often considered individual (Lin & Yang, 2011). Written PF can be advantageous over face-to-face feedback for several reasons. It still provides opportunities for scaffolding and collaborative dialogs, potentially increasing the range of learners' vocabulary, while being less

face-threatening (Chen, 2016) as well as encouraging negotiation and construction of meaning (Liou & Peng, 2009). PF also increases audience awareness, validating the task, and promoting ownership of the text (Lee, 2015; Tsui & Ng, 2000).

Previous research has generally supported the benefits of PF in L2 and FL writing (Diab, 2010; Ekşi, 2012; Lundstrom & Baker, 2009; Mendonça & Johnson, 1994). For example, when Diab (2010) investigated differences between peer-editing and self-editing, she found that the PF group had statistically significantly better compositions, as measured by improved content and organization, following treatment, although both groups had improved. Lundstrom and Baker (2009), who studied the impact of PF on student writing when giving or receiving feedback, found that students in the give-feedback group made more significant gains in their writing (i.e., organization, cohesion, structure, vocabulary, and mechanics) than those in the receive-feedback group. Moreover, students with lower proficiency levels in the give-feedback group improved their writing the most.

Research on PF has mainly explored its effectiveness as compared to instructor feedback (IF) (Lee, 2015; Zhao, 2014). For example, Ekşi's (2012) study examined the writing quality of two groups: one receiving IF and another participating in PF (both giving and receiving) with the exception of the final writing assignment in which the PF group also received IF. Both groups showed writing improvements, indicating the validity of PF. Some studies have also analyzed the effects of PF using CAF measures and have reported that PF leads to significant improvements in complexity, accuracy and fluency in immediate posttests (Soleimani & Rahmanian, 2014), while others have shown significant gains in accuracy and fluency, but not in complexity (Ghahari & Farokhnia, 2018).

2.3 | Collaborative writing versus peer feedback

Although most studies have analyzed the effects of either CW or PF, several studies have also examined both, most often through the creation of wikis or blogs using their L2 (Kessler, 2009; Lee, 2010; Woo et al., 2013), which allows learners to participate in both PF and CW throughout the writing process. Nevertheless, these studies have failed to analyze whether PF or CW might be more beneficial for students' L2 writing skills.

To the best of our knowledge, there is only one study (Bueno-Alastuey & Rodero Albaiceta, 2019), which has compared students' individual texts after experiencing either PF or CW. After a CAF and holistic analysis of the individual pre- and posttest writings of 29 secondary school English language learners, the authors concluded that students who participated in the CW treatment had improved more in complexity and fluency than those who participated in the PF treatment. However, students who participated in the PF treatment showed greater improvement in accuracy and lexical variety than their peers. Both groups showed holistic improvements according to an analytic scale that evaluated the overall quality of the texts. According to this study, both treatments showed positive benefits. Nevertheless, students only had the opportunity to participate in one of the two treatments (PF or CW, not both), and the study lasted just one month, thus, limiting possible developmental changes, or the comparison of both treatments on individual students' texts.

In light of the lack of research investigating CW and PF in the same study and the scarce research on Spanish as a FL using CAF measures, this paper will compare FL Spanish students' initial individual writing assignments and subsequent writings after a CW and a PF treatment. The research questions for the project are:

RQ1: Does the complexity in students' individual writing assignments increase more following CW or PF?
RQ2: Does the accuracy in students' individual writing assignments increase more following CW or PF?
RQ3: Does the fluency in students' individual writing assignments increase more following CW or PF?
RQ4: Does students' overall quality in individual writing assignments improve more following CW or PF?

3 | METHODOLOGY

3.1 | Context

The present study was conducted in two six-credit, intensive, second-year Spanish courses at a large public university in the southwestern United States. The course met 5 days a week for a total of 6 h and used the textbook, *Unidos*, which is designed to be taught following the flipped classroom approach. Outside of this project, no other formal writing assignments were included in the course.

3.2 | Participants

Forty-six consenting students (28 females and 16 males) ranging in age from 19 to 23 years old ($M = 21.5$) participated in this project. They were enrolled in two separate sections of the same intensive Spanish course, forming two intact groups (Group 1 = 22 students, Group 2 = 24 students). All students were native speakers of English, albeit several students were heritage speakers of Spanish. Three students from Group 1 (G1) and two from Group 2 (G2) were excluded from the study due to lack of completion of some part of the project.

3.3 | Instruments

The data collection instruments were three individual writing assignments carried out as scheduled class activities. The first assignment (pretest) took place at the beginning of the course; the second (posttest 1) was collected after the first treatment, either CW or PF; and the third (posttest 2) was collected after switching treatments (see Table 1 for study design). The two treatments, CW or PF, were counterbalanced to prevent the order of treatments from becoming a confounding variable.

Regarding the assignments, and considering that genre has been reported to influence the type of language used and its complexity (Yang et al., 2015), all assignments were narrative essays, but each had a different topic related to the theme and content of the corresponding textbook chapters. The three assignments analyzed in this paper were celebrating a sports victory, throwing a surprise party for someone, and a failed love story (see Appendix III for an example of a prompt). Before each writing assignment, students were given a prewriting task to complete individually as homework. This was the normal procedure to activate vocabulary when they had writing activities and was designed to help them generate ideas for their writing.

TABLE 1 Task schedule

	Week 1 Pretest	Week 2 Treatment	Week 3 Posttest 1	Week 4 Treatment	Week 5 Posttest 2
GROUP 1	Assignment 1 individually	Assignment 2 collaboratively	Assignment 3 individually	Assignment 4 individually + peer feedback	Assignment 5 individually
GROUP 2	Assignment 1 individually	Assignment 2 individually + peer feedback	Assignment 3 individually	Assignment 4 collaboratively	Assignment 5 individually

3.4 | Procedure and tasks

Students completed five writing assignments: three individual (analyzed in this paper), one collaborative, and one with PF (see Table 1).

On the day of the pre-test (Assignment 1), students were given 30 min to produce a narrative essay based on the instructions provided. Students wrote this essay by hand. Both the authors and the course instructors graded the assignment with an analytic scale designed specifically for this course (see Appendix I).

One week later (Week 2), students in G2 (the students who were going to provide PF) participated in a training session led by one of the authors to explain the process of PF. Students were given handouts explaining what to focus on (content, organization, vocabulary, grammar, and mechanics) with examples, and the author modeled providing feedback on those aspects on a student's essay. The next day, students in both groups went to the computer lab. G1 wrote a second narrative essay collaboratively using Google Docs aided by Google Hangouts, which they used to communicate orally with one another. Each member of each pair sat on opposite sides of the computer lab to require them to do the task using Google Hangouts. The process was recorded using Game Bar. G2 wrote a narrative essay individually in Google Docs, shared the document via Google Drive, and provided feedback to a designated classmate. The process of providing feedback was also screen recorded with Game Bar. The pairs in both groups were as homogenous as possible, based on instructors' observations of proficiency level. They were the same throughout the project, as students have been found to give more appropriate feedback to students of a similar level (Storch, 2005). All the sessions lasted 50 min, though the tasks performed during that time period differed between the CW and the PF sessions. Since CW has been reported to take longer (Fernández Dobao, 2012; Storch, 2005), students spent the whole 50 min composing their texts, whereas in the PF sessions students spent 30 min writing and 15 min providing PF. The following week (Week 3), students wrote a second individual 30-min narrative essay (posttest 1) by hand, in class.

One week later (Week 4), students returned to the computer lab and completed the opposite treatments. Before the lab session, students in G1 participated in a training session to explain the process of PF, identical to the session provided to G2 in Week 2. G1 wrote an individual narrative essay in Google Docs and then shared the document for PF, and G2 wrote a narrative essay collaboratively using Google Docs and Google Hangouts. Both processes were also screen recorded. One week later (Week 5), students had 30 min to write a third, individual narrative essay on paper during class (posttest 2), which was assessed by the course instructors and the authors of this article.

3.5 | Data coding

Following similar previous research (Fernández Dobao, 2012; Storch & Wigglesworth, 2007; Storch, 2005; Wigglesworth & Storch, 2009), the current study analyzed the pretest (Assignment 1) and Posttests 1 and 2 (Assignments 3 and 5, respectively). The data analyses involved: (a) quantitative CAF ratings; and (b) a holistic rating using an analytic scale measuring content, organization, grammar, and vocabulary (see Appendix I). To facilitate the analysis, all essays were subsequently typed.

3.5.1 | Analysis of the written texts: CAF

Complexity measures took into consideration syntactic complexity and lexical diversity. *Syntactic complexity* included: (1) the number of clauses per T-unit (C/T), which is a reliable measure, correlating well with other measures of complexity (Foster & Skehan, 1999), and it has been shown to increase at intermediate levels in Spanish (Byrnes et al., 2010); (2) the ratio of dependent clauses to clauses (DC/C), which examines the degree of embedding in the text (Wolfe-Quintero et al., 1998); (3) the number of words per clause, or mean length of clauses (MLC), which has been shown to be a strong predictor of writing quality assessments (Bulté & Housen, 2012); and (4) the number of words per T-unit, or mean length of T-unit (MLTU), a significant predictor of writing quality both in English (Casal & Lee, 2019; Yang et al., 2015) and Spanish (Junco, 1999; Véliz de Vos, 1988).

To calculate complexity measures, T-units and clauses were identified and their quantity computed. T-units were defined as "a main clause plus all subordinate clauses and nonclausal structures attached to or embedded in it" (Hunt, 1970, p. 189). A clause was any unit consisting of a subject (visible or implied) plus a predicate; that is, a construction with a finite or a nonfinite verb as its head (Bulté & Housen, 2012). Clauses were coded as independent and dependent. An independent clause is one that can stand on its own (Richards et al., 1992), while a dependent clause must be used with another clause to form a grammatical sentence (Villarreal & Gil–Sarratea, 2020) and is formed by a finite or a nonfinite verb and at least one additional clause element of the following: subject, object, complement, or adverbial phrase (Foster et al., 2000).

Lexical diversity was calculated using the type-token ratio (TTR), that is, the number of different words divided by the number of total words, and the lexical diversity index (LDI), the number of lexical words divided by the total number of words. Both measures were calculated automatically using online text analysis tools: *Textanalyser,* hosted on the lexicool website (https://www.lexicool.com/textanalyzer.asp?IL=3), for the type-token ratio, and Don Gramaticón, hosted on the *Onomateca* website (http://onomateca.com), to calculate the number of functional and lexical words.

The second measure, *accuracy*, comprised two positive measures—the ratio of error-free T-units to total T-units (EFTU/TU), and the ratio of error-free clauses to total clauses (EFC/C)—and a negative one, the ratio of errors to words (E/W). These three measures were selected to make the results comparable to those of previous research (e.g., Fernández Dobao, 2012; Storch & Wigglesworth, 2007; Storch, 2005; Wigglesworth & Storch, 2009).

Three types of errors, lexical, grammatical, and mechanical, were taken into consideration (e.g., Storch & Wigglesworth, 2007, 2010; Storch, 2007, 2008; Villarreal & Gil–Sarratea, 2020; Wigglesworth & Storch, 2009). If a word had more than one error type, for example, an error in gender and a lack of a graphic accent, it was coded once for each type of error. Extracts 1–3 include examples of errors.[4]

(1) *Grammatical errors* included syntactic errors (missing elements and errors in word order) and morphological errors (errors in gender, use of articles and prepositions, agreement, and use of verb tenses and conjugation).

Example 1. Missing elements: S18_1:
Mis amigas Ø vestieron en ropa similares.
[5][Mis amigas se vestieron en ropa similares.]
[My friends (missing reflexive pronoun) wore similar clothes.]

Example 2. Errors in gender: S12_3:
Pero, una día, después de ocho meses,
*[Pero, un día, después de ocho meses,]
[But, one (feminine) day (masculine), eight months later,]

(2) *Lexical errors* covered issues in word choice (words from other language or borrowings), including incorrect words within the right category.

Example 3. Wrong choice of preposition:
S7_1:*Mi mama y mi hermana asistieron para la fiesta con muchas comida*
*[Mi mama y mi hermana asistieron a la fiesta con muchas comida]
[My mother and my sister went to the party with a lot of food]

Example 4. Transfer: S13_2:*En la sobre de "rojo A" es la sobre la puritans en primero Estados Unidos.*
*[En la [película] sobre de "rojo A" es la sobre la puritanos en primero Estados Unidos.]
[In the [movie] about the "scarlet letter", [which] is about puritans in the early United States.]

(3) *Mechanical errors* include spelling, punctuation, capitalization, and lack of graphic accents in Spanish.

Example 5. Spelling errors: S13_3:
de el
*[del]
[from]

Example 6. Punctuation errors: S22_1:*ØEl equipo ganó el partido por más de triente puntos!*
*[¡El equipo ganó el partido por más de triente puntos!]
(Missing initial exclamation mark)
[The team won the match by over thirty points!]

Fluency was measured by the total number of words, T-units, and clauses per text (Rezazadeh et al., 2011; Wigglesworth & Storch, 2009).

3.5.2 | Analysis of the written texts: Holistic rating

In addition to the quantitative analysis, a holistic analysis was carried out to have a qualitative perception of the quality of the texts. The texts were scored using a 4-factor 6-point analytic

scale (see Appendix I), where 1 represented the lowest mark and 6 the highest. The four factors rated in the scale were:

i. Content, which evaluated idea development and the use of supporting details;

ii. Structure and organization, which considered whether the texts contained all the content required in the instructions: intro, body, and conclusion, and if the information was organized in a logical sequence and in three paragraphs as instructed;

iii. Vocabulary, which evaluated the appropriateness, range, and variety of the vocabulary used; and

iv. Grammar, which assessed the accuracy, appropriateness, variety, and complexity of the grammar used.

3.6 | Data analysis procedure

First, one student's essay was coded by the three authors of this paper to determine coding categories: T-units, clauses, dependent and independent clauses, errors and types of errors, lexical and functional elements, and the number of words were identified. Grades were also provided for the qualitative holistic ratings of content, organization, grammar, and vocabulary for that essay. Once an agreement was reached, seven essays were coded and assessed by the three authors independently. Interrater reliability using Krippendorff's α (alpha) reliability estimate (Hayes & Krippendorff, 2007) was .8931. The remaining differences were discussed until complete agreement was reached.

In the second round of coding, the researchers coded and assessed two more essays independently (Krippendorff's $\alpha = .8992$), and discussed the results and the differences until total agreement was reached. Two additional essays were coded independently (Krippendorff's α interrater reliability was .9381) and again differences were discussed until agreement was reached. After those three rounds, each of the authors coded 40 or 41 more essays (91% of the total) independently.

Descriptive inferential statistics were calculated in SPSS. Since, according to a Sapiro–Wilk's test, the data were not normally distributed, nonparametric tests were used. Mann–Whitney *U* tests for independent samples were run to test for statistically significant differences between Groups 1 and 2, and Wilcoxon signed-rank tests for related samples were calculated to search for statistically significant differences within each group. Effect sizes using Cohen's (1988) r —0.1 represents a small effect size, 0.3 medium, and 0.5 large—were calculated for significant results.

4 | RESULTS

4.1 | Improvement of students' writing skills: CAF measures

4.1.1 | Complexity

When comparing syntactic complexity (see Table 2) between groups, initially G2 had higher scores than G1 in C/T (+0.11) and DC/C (+0.06), but lower in MLC (−0.95) and MLTU (−0.41).

TABLE 2 Results of complexity measures

		Group 1			Group 2		
		Pretest	Postcollaborative writing	Postpeer feedback	Pretest	Postpeer feedback	Postcollaborative writing
Syntactic complexity	C/T	1.12 (0.06)	1.15 (0.11)	1.28 (0.14)	1.23 (0.21)	1.29 (0.25)	1.32 (0.26)
	DC/C	0.10 (0.05)	0.16 (0.19)	0.21 (0.08)	0.16 (0.14)	0.23 (0.1)	0.3 (0.25)
	MLC	7.3 (1.8)	6.2 (0.68)	5.5 (0.77)	6.35 (0.94)	6.30 (1.26)	5.9 (0.9)
	MLTU	8.18 (2.1)	7.13 (0.9)	7.08 (1.17)	7.77 (1.46)	8.04 (1.7)	7.75 (1.55)
Lexical diversity	TTR	0.56 (0.05)	0.55 (0.06)	0.60 (0.08)	0.59 (0.08)	0.58 (0.07)	0.60 (0.05)
	LDI	0.52 (0.03)	0.51 (0.03)	0.54 (0.04)	0.53 (0.03)	0.51 (0.03)	0.54 (0.06)

Note: Mean (*SD*).

Abbreviations: C/T, clauses to T-units; DC/C, ratio of dependent clauses to clauses; LDI, Lexical Diversity Index; MLC, mean length of clauses; MLTU, mean length of T-unit; SD, standard deviation; TTR, type token ratio.

None of these differences between groups were statistically significant in the pretest. After posttest 1 (G1-CW, G2-PF), G2 increased their scores more than G1 in all four measures. That is, G2's scores were significantly higher and with a medium effect size in C/T (+0.14; $U = 107.500$, $z = -2.655$, $p = .008$, $r = 0.41$) and DC/C (+0.07; $U = 105.500$, $z = -2.707$, $p = .007$, $r = 0.42$), and higher, but not significantly, in MLC (+0.1) and MLTU (+0.91). After posttest 2 (G1-PF, G2-CW), G2 still had higher scores than G1 in all four parameters (C/T + 0.04, DC/C + 0.09, MLC + 0.04, MLTU + 0.67). However, none of the differences between groups were statistically significant after posttest 2; that is, the scores of the two groups became more similar. G2 improved both after PF and after CW, but their results became significantly higher than G1's results and had a medium effect size only after PF in C/T and DC/C. G1's results also improved after CW, but improved more after PF, as the significantly lower scores (when compared with G2) on posttest 1 (CW) on C/T and DC/C became nonsignificant on posttest 2 (PF).

Considering the performance within groups after each treatment, G1 increased their scores, after CW from the pretest to posttest 1, in the ratios of C/T (+0.03) and DC/C (+0.06), but decreased their scores significantly in MLC (−1.1; $z = -2.495$, $p = .013$, $r = 0.4$) and MLTU (−1.05; $z = -2.334$, $p = .020$, $r = 0.38$). The same trend occurred from posttest 1 (following the CW treatment) to posttest 2 (following PF treatment) within G1. There was a statistically significant increase with a large effect size in C/T (+0.13; $z = -3.260$, $p = .001$, $r = 0.53$) and with a medium effect size in DC/C (+0.05; $z = -2.374$, $p = .018$, $r = 0.38$), and a decrease in MLC (−0.7) and MLTU (−0.05). Only the former was statistically significant ($z = -2.736$, $p = .006$, $r = 0.44$) with a medium effect size. The four measures reached statistical significance with large effect sizes (except MLTU with a medium effect size) for G1 from pre-test to posttest 2 (PF): C/T ($z = -3.179$, $p = .001$, $r = 0.52$), DC/C ($z = -3.179$, $p = .001$, $r = 0.52$), MLC ($z = -3.099$, $p = 0.002$, $r = 0.50$), and MLTU ($z = -2.052$, $p = .04$, $r = 0.33$).

The results for the indices of syntactic complexity of G2 followed similar tendencies. G2's scores increased nonsignificantly, from pretest to posttest 1 (following PF), in the ratio of C/T (+0.04), of DC/C (+0.07) and, contrary to what happened in G1, in MLTU (+0.27). As in G1, MLC decreased (−0.05) for G2 from pretest to posttest 1. Only the increase in the DC/C was statistically significant ($z = -2.172$, $p = .03$, $r = 0.33$) with a medium effect size. The results from posttest 1 (PF) to posttest 2 (following CW) were similar to G1's results. G2 experienced an increase in the ratio of C/T (+0.03), and DC/C (+0.07), and a decrease in MLC (−0.01) and MLTU (−0.29). None of the differences from posttest 1 (PF) to posttest 2 (CW) in G2 reached statistical significance. Regarding the comparison between pretest and posttest 2 results, only the MLC ($z = -2.289$, $p = .022$, $r = 0.34$) reached statistical significance from pre-test to posttest 2 (CW).

With regard to lexical density (see Table 2), comparisons between groups did not reveal statistical differences: scores were quite similar across tasks and groups and were slightly higher in posttest 2 irrespective of the treatment. Both groups obtained the highest lexical diversity rates in the second task, 0.60 in TTR and 0.54 in LRI. There were no statistical differences between groups in any measures of lexical diversity.

Regarding results within groups after the first treatment (G1-CW, G2-PF), both groups decreased their scores slightly in both measures, TTR (both −0.01) and LDI (G1 −0.01, G2 −0.02), from pretest to posttest 1. The decrease in LDI was statistically significant in G2 ($z = -2.289$, $p = 0.022$, $r = 0.34$) with a medium effect size. On the contrary, after the second treatment (G1-PF, G2-CW), both groups increased their scores in both measures: TTR (G1 + 0.05, G2 + 0.02) and LDI (both + 0.03). However, only the scores in G1 reached statistical

TABLE 3 Results of accuracy measures

	Group 1			Group 2		
	Pretest	Postcollaborative writing	Postpeer feedback	Pretest	Postpeer feedback	Postcollaborative writing
EFTU/TU	0.15 (0.10)	0.17 (0.12)	0.24 (0.18)	0.22 (0.10)	0.22 (0.15)	0.25 (0.17)
EFC/C	0.14 (0.1)	0.19 (0.13)	0.28 (0.17)	0.27 (0.13)	0.29 (0.19)	0.31 (0.18)
E/W	0.28 (0.1)	0.30 (0.14)	0.27 (0.11)	0.22 (0.08)	0.21 (0.07)	0.21 (0.09)

Note: Mean (*SD*). Higher values indicate more errors.

Abbreviations: EFTU/TU, ratio of error-free T-Units to total T-Units; EFC/C, ratio of error-free clauses to total clauses; E/W, ratio of errors to words, which is a negative index.

significance in TTR from pretest to posttest 2 (PF) ($z = -2.178$, $p = .029$, $r = 0.35$) and from posttest 1 (CW) to posttest 2 (PF) ($z = -2.93$, $p = .003$, $r = 0.48$). Both had medium effect sizes.

Overall, as regards syntactic complexity, G1 performed worse after CW and better after PF, while G2 performed better after both treatments. Nevertheless, only G1's results reached statistical significance. Regarding lexical diversity, both groups increased their scores after the second treatment irrespective of whether it was CW or PF. Although initially the scores of G2 were slightly higher, G1's scores increased significantly more from pre-test to posttest 2, and from posttest 1 to posttest 2.

4.1.2 | Accuracy

At pre-test, the analysis of accuracy measures between the groups (see Table 3) revealed that G1's scores were significantly less accurate with medium effect sizes than G2's in all measures: EFTU/TU (−0.07; $U = 118.00$, $z = -2.380$, $p = .017$, $r = 0.37$), EFC/C (−0.13; $U = 104.00$, $z = -2.747$, $p = .006$, $r = 0.43$), and EW (+0.06; $U = 128.00$, $z = -2.118$, $p = .034$, $r = 0.33$). No significant differences were obtained when accuracy results were contrasted in the pretest and posttest 1. After the first treatment (CW), G1 increased their scores in EFTU/TU (+0.02) and EFC/C (+0.05), but had more errors per words (EW) (+0.02); while G2 (PF) maintained the same score in EFTU/TU, and increased them in EFC/C (+0.02), and in EW (−0.01). Although the scores after the first treatment were higher for G2, their improvement was less salient than in G1. In fact, the initial significant differences between both groups disappeared after the first treatment, when none of the differences were statistically significant from pre-test to posttest 1. This indicates that G1's scores increased more when compared with G2's, even if G2's scores remained higher overall.

After the second treatment (G1-PF, G2-CW), from posttest 1 to 2, both groups increased their scores but nonsignificantly in the first two measures (EFTU/TU: G1 + 0.07, G2 + 0.03 and EFC/C: G1 + 0.09, G2 + 0.02). However, only G1 received higher scores in EW (−0.03) while the ratio remained the same for G2. None of the differences between the groups reached statistical significance in posttest 2. However, considering the scores within groups, G1's scores increased significantly from posttest 1 to posttest 2 in EFC/C ($z = -2.069$, $p = .039$, $r = 0.34$), while none of the increases were statistically significant within G2.

In sum, regarding accuracy (see Table 3), G1 improved more than G2. Although G2 had statistically significant higher scores on the pre-test, G1 increased their scores more on both

TABLE 4 Results of fluency

	Group 1			Group 2		
	Pretest	Postcollaborative writing	Postpeer feedback	Pretest	Postpeer feedback	Postcollaborative writing
Words	174.41 (53.4)	149.16 (35.6)	149.37 (43)	134.36 (37.6)	140.95 (29.9)	139.45 (38.2)
T-Units	22 (7.2)	21.37 (6.3)	21.58 (6.8)	17.45 (4.05)	18 (4.5)	18.64 (6.3)
Clauses	24.63 (8)	24.44 (6.7)	27.37 (7.5)	21.32 (5.9)	23 (6.5)	24.09 (7.3)

Note: Mean (*SD*).

posttests and the scores between groups became more similar. Although the results of both treatments were positive for accuracy, more gains were observed when the first treatment was CW and the second PF.

4.1.3 | Fluency

Regarding fluency (see Table 4), G1 wrote significantly more words (174.41 vs. 134.36; $U = 114.5$, $p = .013$) and more T-units (22 vs. 17.45) and clauses (24.63 vs. 21.32) than G2 in the pretest, in posttest 1 (149.16 vs. 140.95; 21.37 vs. 18; 24.44 vs. 23, respectively) and in posttest 2 (149.37 vs. 139.45; 21.58 vs. 18.64; 27.37 vs. 24.09, respectively). Only the number of words in the pretest reached statistical significance.

Comparing performance within groups, G1 wrote significantly fewer words (24.14; $Z = -2.013$, $p = .044$) and fewer T-units (−0.63) and clauses (−0.19) in posttest 1 (CW), and fewer words (−0.97), but more T-units (+0.21) and clauses (+2.93) in posttest 2 after PF. On the contrary, G2 wrote more words (+13.21), T-units (+0.55), and clauses (+1.68) in posttest 1 after PF, and fewer words (−1.17), but more T-units (+0.64) and clauses (+1.09) after CW. Consequently, both groups wrote more T-units and clauses after PF, but, unlike G1, G2 also increased their quantity after CW. Only the number of words from pre-test to post-test1 in G1 reached statistical significance.

4.2 | Improvement of EFL students' writing skills: Holistic ratings

Holistic measures did not yield significant group differences (Table 5). At pretest, G1 had a higher holistic score than G2 (14.72 vs. 13.18), but G2 outperformed (nonsignificantly) G1 (16.27 vs. 14.17) after the first treatment (G1-CW, G2-PF). After the second treatment (G1-PF, G2-CW), the reverse happened: G1 increased their scores nonsignificantly and had higher results than G2 (16.11 vs. 15.64).

Comparing the performance within each group, G1's holistic scores decreased from 14.79 to 14.17 from pretest to posttest 1 (CW) and increased significantly with a medium-size effect after PF to 16.11 ($z = -2.335$, $p = .020$, $r = 0.38$). A similar pattern was observed in G2. The group improved after PF from 13.18 to 16.27 and this difference was statistically significant with a medium-size effect ($z = -2.355$, $p = .019$, $r = 0.35$), while their scores decreased after CW to 15.64. This difference did not reach statistical significance, although the difference from pretest to posttest 2 remained statistically significant ($z = -2.066$, $p = .039$, $r = 0.31$).

TABLE 5 Results of holistic measures (max = 24 points)

	Group 1			Group 2		
	Pretest	Postcollaborative writing	Postpeer feedback	Pretest	Postpeer feedback	Postcollaborative writing
Total	14.72 (4.2)	14.17 (4.3)	16.11 (3.7)	13.18 (5.1)	16.27 (3.9)	15.64 (4)

Note: Mean (*SD*).

In summary, regarding holistic ratings, both groups performed better after PF and worse after CW irrespective of whether it was the first or the second treatment.

5 | DISCUSSION

The main goal of this study was to compare the effect of two writing interventions on subsequent individual writing assignments. Regarding the first research question, which aimed to analyze differences in complexity in individually written essays following PF and CW treatments, both G1 and G2 performed better following PF. Additionally, G2 also performed better following CW, which they completed as their second treatment after having experienced PF. In fact, G1's scores in both MLTU and MLC continued to decrease even after PF, while G2 only decreased both scores after CW. Our findings seem to support previous studies that had already reported the lack of significant differences in syntactic complexity between CW and individual texts (Fernández Dobao, 2012; McDonough et al., 2018; Strobl, 2014; Wigglesworth & Storch, 2009) and even suggested improved syntactic complexity when writing individually (Bueno-Alastuey & Martínez de Lizarrondo, 2017; McDonough et al., 2018). These results also confirm previous results on the effect of CW on subsequent individual writing (Chen, 2019; Elabdali, 2021) and point to the fact that any increase in syntactic complexity reported by previous studies comparing CW and individual texts (Storch, 2005) does not seem to be retained when students write individually afterward, as both groups obtained lower scores after CW. These findings may point to the difficulty of noticing syntactic structures (e.g., dependent clauses) when working collaboratively. The improvement of both groups following PF also supports previous studies that analyze the effects of PF (Soleimani & Rahmanian, 2014), but contradicts Bueno-Alastuey and Rodero Albaiceta (2019), who compared CW and PF in an EFL context. These results indicate that even if PF does not appear to aid the development of syntactic complexity in EFL, this technique might aid in its development in Spanish even at low levels of proficiency. Both groups decreased in MLC and MLTU while they increased their scores in subordination indices. This is consistent with previous research indicating that an increase in subordination implies improvement at intermediate levels (Byrnes et al., 2010).

With reference to the order of the treatments, two of the measures (C/T & DC/C) increased significantly and with large effect sizes from pretest to posttest 2 only in G1, while none increased significantly in G2. Consequently, our findings seem to imply that experiencing CW followed by PF is conducive to significantly better syntactic complexity development, although experiencing PF followed by CW is also conducive to gains (even if not statistically significant).

Concerning lexical diversity, both groups performed better after the second treatment irrespective of whether it was CW or PF. Supporting the results of previous research (Bueno-Alastuey & Rodero Albaiceta, 2019), both groups' scores decreased very slightly after the first

treatment. However, both groups improved after the second treatment. Consequently, it seems that more than one treatment is needed to observe positive effects in lexical diversity, and therefore, writing using different treatments and various tasks should be carried out in courses to provoke lexical advancement.

The second research question aimed to examine whether the accuracy of students' individual writing scores increased more following CW or PF. Our findings reveal that after CW students' scores improved irrespective of whether it was the first or the second treatment supporting previous studies indicating better results in subsequent writings after CW (Chen, 2019). However, PF only produced noticeable improvements when it was the second treatment following CW, which contradicts Bueno-Alastuey and Rodero Albaiceta's (2019), study, which showed improvements after PF, but not after CW, and points to the combination of CW followed by PF as the best combination to increase the accuracy of writing essays. As suggested by previous researchers, these findings seem to support the idea that collaborative tasks help students notice and discuss linguistic forms, so they are able to produce more accurate texts (Nassaji & Tian, 2010; Storch & Wigglesworth, 2007; Tian, 2011). In the current study, this higher accuracy remains in subsequent writings, which may mean that the combination of both treatments may have a longer-term impact on students' individual texts. A longer study may help confirm the validity of these results.

With respect to the third research question, which analyzed fluency in students' individual writing after CW and PF, and contrary to the results reported by Bueno-Alastuey and Rodero Albaiceta (2019), both groups increased their fluency scores after PF, supporting previous research (Chen, 2019; Ghahari & Farokhnia, 2018; Soleimani & Rahmanian, 2014), while only G2 increased them after CW, which was their second treatment. The combination of PF followed by CW produced better results, although not statistically significant, for fluency.

Finally, regarding the fourth question, which queried whether students improved the holistic scores of their texts more following CW or PF, our findings point to better outcomes immediately following PF for both G1 and G2. Students performed better after PF irrespective of whether it was the first or the second treatment. These results support previous research indicating improvements after PF (Diab, 2010; Ekşi, 2012). However, they contradict Bikowski and Vithanage's (2016) and Chen's (2019) studies reporting significant increases after CW, and Bueno-Alastuey and Rodero Albaiceta's (2019) study, whose secondary school students performed slightly worse after PF than after CW. These differences could be due to the different learning contexts (secondary vs. university) and the different FL (English vs. Spanish). It could also be the case that students in this study were more comfortable providing and giving feedback and working independently and autonomously due to the flipped nature of their course, therefore, enabling them to integrate their knowledge from PF sessions more than the students in the previous study. Furthermore, students in the current study were older and cognitively more prepared to integrate new knowledge through PF. They also may have had more experience with and paid more attention to their peers' feedback. Giving feedback might also have had a positive influence in their results as mentioned by previous research (Lundstrom & Baker, 2009; Rouhi et al., 2020), which showed that giving feedback produced more significant improvements than just receiving it. This seems to be also the case even with low-level students, an under-researched cohort.

While working collaboratively, students might have focused more on content and less on correcting form or they might have been distracted while the other member of the pair was writing. It might also be the case that, with PF, there is personal individual responsibility for error identification, whereas in CW there is (presumably) joint responsibility, so students might

pay less attention on the assumption that their partner will catch any mistakes. Another explanation might be that students tend to retain information better when they provide and receive feedback than when they work collaboratively. In this case, the need to focus on several writing processes might limit their attentional resources (Skehan, 2009) more than when concentrating solely on the revision and feedback provision. They might experience a higher cognitive load during CW, as they have to focus on both creating a text, discussing, and correcting errors and choices, than during PF when they only have to focus on error correction and on providing alternatives since the text is already written.

Another interesting finding is that students' scores seemed to decrease after CW if it was the first treatment but not when it was the second treatment. Thus, we can tentatively suggest that treatment order might have an effect on the results. It seems that PF followed by CW produces better results regarding general quality and fluency, while CW followed by PF is more beneficial for syntactic complexity and accuracy.

6 | LIMITATIONS AND FURTHER RESEARCH

As in any research study, there were limitations that may have skewed some of the results and can be addressed in future research. The individual writing tasks (pre-tests and post-tests), for instance, were carried out on paper while the treatments were computer-based. That might have affected the results, as many word processing programs, such as Google Docs, include grammar and spelling support and have the potential to suggest grammar, style, and spelling changes. These suggestions and automatic corrections might have aided students to notice mistakes that they would not have focused on when writing on paper. Future research should explore whether carrying out all treatments on computers might influence the results.

The lack of a control group might have also prevented a clearer picture of CW's and PF's potential benefits as well as the treatment order benefits. The question that arises is whether students improved their writing overall because of a task repetition effect, which helped students become familiar with the genre of narration, or because of the treatments themselves. This ambiguity in the results could have been counteracted by including a control group.

Another limitation was the lack of training on how to write collaboratively—unlike the PF training students received—which may account for less successful results in the first treatment (CW) for G1, when students might not have known each other well and were not yet familiar with the writing assignments. Furthermore, the impact of the PF training on the feedback given might have affected the results.

Task and genre effects are also important to investigate in the future. In this study, only one genre and one type of task were studied. Future research could investigate the effect of task and genre when these treatments are used.

Another issue is that there are no insights into the interactions and conversations that took place when working collaboratively or when carrying out PF. Further studies could explore what participants do and talk about when they write collaboratively and when they give PF. This would provide insights about what areas students focus on and if these interactions and areas of focus have any impact on subsequent individual writing assignments. These data would also shed light on pair dynamics and may provide information about how individual characteristics, such as being a heritage learner, affect group interactions and subsequent individual texts. Additionally, the analysis of LREs, type of interactions, and type of feedback

(e.g., oral vs. written feedback) will offer a better understanding of the impact of diverse treatments.

Finally, the results could be more conclusive if the students had received more treatments and the study had been more longitudinal in nature, as that would provide information about whether treatment order (better performance after PF followed by CW) was sustained.

7 | CONCLUSION

The findings of this study point to the beneficial effect of both practices to improve different aspects of writing and to the even more beneficial impact of combining them. Even though there has been a plethora of research on the benefits of CW practices, mainly on ESL or EFL, students of Spanish as an FL seem to benefit the most from PF practices, most likely because of the focus on form PF promotes. Our results seem to advocate for the use of combined CW and PF for syntactic complexity and accuracy, and PF followed by CW to improve general quality and fluency.

There seems to be a positive effect on individual writing, especially when PF is conducted before CW. Combining CW and PF seems to be beneficial for FL courses for several reasons: PF and CW may have a complementary effect when students are working with FL writing at low proficiency levels (first and second years of language study); training students and the order of treatment may also need to be considered until students feel comfortable working with their peers, thus starting with PF may be more effective because students are more familiar with this activity than perhaps composing collaboratively; and, finally, certain measures such as accuracy and lexical diversity may benefit from these treatments more rapidly than syntactic complexity due to a need for a higher proficiency and pragmatic threshold.

The final suggestion for language educators is that exposing students to different writing approaches such as CW, PF, and individual writing provide students with a range of writing practices and approaches that may be missing when students are asked to only work individually.

ACKNOWLEDGMENTS

The first author would like to thank the Public University of Navarre for providing financial support to carry out this study and to publish it open access, Texas Tech University for hosting her and providing the facilities and means to do the experiment. The three authors would like to thank Chris Vasquez-Wright, PhD, the director of the language lab, and research center at Texas Tech University for his help with the collection and storage of data, Izaskun Villarreal, PhD, for her valuable comments on previous versions of this article, and the reviewers for their useful feedback to improve the article.

ORCID

Mª Camino Bueno-Alastuey http://orcid.org/0000-0001-7027-5382

ENDNOTES

[1] Heritage language: In this study, it refers to the teaching of Spanish in the United States to students who grew up hearing Spanish at home. These students may speak the Spanish language or not (receptive students).

[2] Collaborative writing: Two or more students composing a single text together.

[3] Peer-feedback: Students providing written or oral feedback to one another.

[4] Only the error referred to within the specific category of errors appears underlined even if there are other errors in the same sentence.

[5] Correct version of the error in Spanish.

REFERENCES

Al-Jamhoor, M. M. (2005). *Connecting Arabs and Americans online to promote peace and to increase cultural awareness: A descriptive study about Arab EFL learners' perceptions, practices, behaviors and attitudes towards computer-supported collaborative writing strategies and technologies* (Publication No. 3164696.) [Doctoral dissertation, Indiana University of Pennsylvania]. ProQuest.

Alyousef, H. S., & Picard, M. Y. (2011). Cooperative or collaborative literacy practices: Mapping metadiscourse in a business students' wiki group project. *Australasian Journal of Educational Technology*, *27*(3), 463–480.

Arnold, N., Ducate, L., & Kost, C. (2009). Collaborative writing in wikis: Insights from culture project in German class. In L. Lomicka, & G. Lord (Eds.), *The next generation: Social networking and online collaboration in foreign language learning* (pp. 115–144). Texas State University.

Aydin, Z., & Yildiz, S. (2014). Using wikis to promote collaborative EFL writing. *Language Learning & Technology*, *18*(1), 160–180.

Bikowski, D., & Vithanage, R. (2016). Effects of web-based collaborative writing on individual L2 writing development. *Language Learning & Technology*, *20*(1), 79–99.

Biria, R., & Jafari, S. (2013). The Impact of collaborative writing on the writing fluency of Iranian EFL Learners. *Journal of Language Teaching & Research*, *4*(1), 164–175.

Bosley, D. S. (1989). *A national study of the uses of collaborative writing in business communication courses among members of the ABC* [Unpublished doctoral dissertation]. Illinois State University.

Bueno-Alastuey, M. C., & Martínez de Lizarrondo, P. (2017). Collaborative writing in the EFL Secondary Education classroom: Comparing triad, pair and individual work. *Huarte de San Juan. Filología y Didáctica de la Lengua*, *17*, 254–275.

Bueno-Alastuey, M. C., & Rodero Albaiceta, S. (2019). The effects of using collaborative writing vs. peer review treatments on subsequent individual writings. *Huarte de San Juan. Filología y Didáctica de la Lengua*, *19*, 32–61.

Bulté, B., & Housen, A. (2012). Defining and operationalising L2 complexity. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA* (Vol. 27, pp. 23–46). John Benjamins Publishing.

Byrnes, H., Maxim, H. H., & Norris, J. M. (2010). Realizing advanced foreign language writing development in collegiate education: Curricular design, pedagogy, assessment. *The Modern Language Journal*, *94*, s1–202.

Casals, J. E., & Lee, J. J. (2019). Syntactic complexity and writing quality in assessed first-year L2 writing. *Journal of Second Language Writing*, *44*, 51–62.

Chen, T. (2016). Technology-supported peer feedback in ESL/EFL writing classes: A research synthesis. *Computer Assisted Language Learning*, *29*(2), 365–397.

Chen, W. (2019). An exploratory study on the role of L2 collaborative writing on learners' subsequent individually composed texts. *Asia-Pacific Education Research*, *28*, 563–573.

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

Diab, N. M. (2010). Effects of peer-versus self-editing on students' revision of language errors in revised drafts. *System*, *38*(1), 85–95.

Ekşi, G. (2012). Peer review versus teacher feedback in process writing: How effective? *IJAES*, *13*(1), 33–48.

Elabdali, R. (2021). Are two heads really better than one? A meta-analysis of the L2 learning benefits of collaborative writing. *Journal of Second Language Writing*, *52*, 1–16.

Elola, I. (2017). Writing in Spanish as a second and heritage language: Past, present, and future. *Hispania*, *100*(5), 119–124.

Fernández Dobao, A. (2012). Collaborative writing tasks in the L2 classroom: Comparing group, pair, and individual work. *Journal of Second Language Writing*, *21*, 40–58.

Fernández Dobao, A., & Blum, A. (2013). Collaborative writing in pairs and small groups: Learners' attitudes and perceptions. *System*, *41*(2), 365–378.

Foster, P., & Skehan, P. (1999). The influence of source of planning and focus of planning on task-based performance. *Language Teaching Research*, *3*(3), 215–247.

Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. *Applied Linguistics*, *21*(3), 354–375.

García Mayo, M. (2007). *Investigating tasks in formal language learning*. Multilingual Matters.

Ghahari, S., & Farokhnia, F. (2018). Peer versus teacher assessment: Implications for CAF triad language ability and critical reflections. *International Journal of School & Educational Psychology*, *6*(2), 124–137.

Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, *1*, 77–89.

Hunt, K. W. (1970). Syntactic maturity in schoolchildren and adults. *Monographs of the Society for Research in Child Development*, *35*, iii67.

Junco, M. (1999). La evaluación del discurso narrativo escrito en el nivel elemental de enseñanza. *Milenio*, *3*, 213–244.

Kessler, G. (2009). Student-initiated attention to form in wiki-based collaborative writing. *Language Learning & Technology*, *13*(1), 79–95.

Kuiken, F., & Vedder, I. (2002). Collaborative writing in L2: The effect of group interaction on text quality. In S. Ransdell, & M. L. Barbier (Eds.), *New directions for research in L2 writing* (Vol. 31, pp. 169–188). Springer Science and Business Media.

Lee, L. (2010). Exploring wiki-mediated collaborative writing: A case study in an elementary Spanish course. *CALICO Journal*, *27*(2), 260–272.

Lee, M. (2015). Peer feedback in second language writing: Investigating junior secondary students' perspectives on inter-feedback and intra-feedback. *System*, *55*, 1–10. https://doi.org/10.1016/j.system.2015.08.003

Limbu, L., & Markauskaite, L. (2015). How do learners experience joint writing: University students' conceptions of online collaborative writing tasks and environments. *Computers & Education*, *82*, 393–408.

Lin, W. C., & Yang, S. C. (2011). Exploring students' perceptions of integrating Wiki technology and peer feedback into English writing courses. *English Teaching: Practice and Critique*, *10*, 88–103.

Liou, H. C., & Peng, Z. Y. (2009). Training effects on computer-mediated peer review. *System*, *37*(3), 514–525.

Long, M. H. (1996). The role of the linguistic environment in second language acquisition. In W. C. Ritchie, & T. K. Bhatia (Eds.), *Handbook of language acquisition* (Vol. 2, pp. 413–468). Academic.

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, *15*(4), 474–496.

Lundstrom, K., & Baker, W. (2009). To give is better than to receive: The benefits of peer review to the reviewer's own writing. *Journal of Second Language Writing*, *18*(1), 30–43.

McDonough, K., De Vleeschauwer, J., & Crawford, W. (2018). Comparing the quality of collaborative writing, collaborative prewriting, and individual texts in a Thai EFL context. *System*, *74*, 109–120.

Mendonça, C. O., & Johnson, K. E. (1994). Peer review negotiations: Revision activities in ESL writing instruction. *TESOL Quarterly*, *28*(4), 745–769.

Nassaji, H., & Tian, J. (2010). Collaborative and individual output tasks and their effects on learning English phrasal verbs. *Language Teaching Research*, *14*(4), 397–419.

Olovson, B. (2018). *Are two heads better than one? A process and product analysis of collaborative writing in the Spanish as a foreign language classroom*. (Publication No. 9983776828202771) [Doctoral dissertation, University of Iowa] Iowa Research Online.

Oskoz, A., & Elola, I. (2011). *Meeting at the wiki: The new arena for collaborative writing in foreign language courses. In Web 2.0-based e-learning: Applying social informatics for tertiary teaching* (pp. 209-227). IGI Global.

Oskoz, A., & Elola, I. (2013). Beyond the FL writing classroom: Social tools at work. In N. Estévez Fuerte, & B. Clavel Arroitia (Eds.), *Adquisición de segundas lenguas en el marco del nuevo milenio* (pp. 211–228). Universitat de València.

Oskoz, A., & Elola, I. (2020). *L2 digital writing literacies*. Equinox Press.

Oskoz, A., & Elola, I. (2012). Understanding the impact of social tools in the FL classroom: Activity theory at work. In G. Kessler, A. Oskoz, & I. Elola (Eds.), *Technology across writing contexts and tasks* (pp. 131–153). CALICO.

Oskoz, A., & Elola, I. (2014). Promoting FL collaborative writing through the use of Web 2.0 tools. In M. González-Lloret, & L. Ortega (Eds.), *Technology and tasks: Exploring technology-mediated TBLT* (pp. 115–147). John Benjamins.

Polio, C., & Shea, M. C. (2014). An investigation into current measures of linguistic accuracy in second language writing research. *Journal of Second Language Writing*, *26*, 10–27.

Rezazadeh, M., Tavakoli, M., & Rasekh, A. E. (2011). The role of task type in foreign language written production: Focusing on fluency, complexity, and accuracy. *International Education Studies*, *4*(2), 169–176.

Richards, J. C., Platt, J., & Platt, H. (1992). *Dictionary of language teaching and applied linguistics*. Longman

Rouhi, A., Dibah, M., & Mohebbi, H. (2020). Assessing the effect of giving and receiving written corrective feedback on improving L2 writing accuracy: does giving and receiving feedback have fair mutual benefit? *Asian-Pacific Journal of Second and Foreign Language Education*, *5*(1), 1–13.

Shehadeh, A. (2011). Effects and student perceptions of collaborative writing in L2. *Journal of Second Language Writing*, *20*(4), 286–305.

Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics*, *30*(4), 510–532.

Soleimani, H., & Rahmanian, M. (2014). Self-, peer-, and teacher-assessments in writing improvement: A study of complexity, accuracy, and fluency. *Research in Applied Linguistics*, *5*(2), 128–148.

Storch, N. (2005). Collaborative writing: Product, process, and students' reflections. *Journal of Second Language Writing*, *14*(3), 153–173.

Storch, N. (2007). Investigating the merits of pair work on a text editing task in ESL classes. *Language Teaching Research*, *11*(2), 143–159.

Storch, N. (2008). Metatalk in a pair work activity: Level of engagement and implications for language development. *Language Awareness*, *17*(2), 95–114.

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

Storch, N., & Wigglesworth, G. (2007). Writing tasks: The effects of collaboration. In M.P. García Mayo (Ed.), *Investigating tasks in formal language learning* (pp. 157–177). Multilingual Matters.

Strobl, C. (2014). Affordances of web 2.0 technologies for collaborative advanced writing in a foreign language. *Calico Journal*, *31*(1), 1–18.

Suzuki, W., & Storch, N. (2020). *Languaging in language learning and teaching: A collection of empirical studies*. John Benjamins.

Swain, M. (2010). 'Talking-it-through': Languaging as a source of learning. In R. Batstone (Ed.), *Sociocognitive perspectives on language use and language learning* (pp. 112–130). Oxford University Press.

Swain, M. (2000). The output hypothesis and beyond: Mediating acquisition through collaborative dialogue. In J. P. Lantolf (Ed.), *Sociocultural theory and second language learning* (pp. 97–114). Oxford University Press.

Tian, J. (2011). *The Effects of a peer editing versus co-writing on writing in Chinese-as-a-foreign language*. [Doctoral dissertation, University of Victoria]. UVicSpace.

Tsui, A. B., & Ng, M. (2000). Do secondary L2 writers benefit from peer comments? *Journal of Second Language Writing*, *9*(2), 147–170.

Véliz de Vos, M. (1988). Evaluación de la madurez sintáctica en el discurso escrito. *Revista de Lingüística Teórica Aplicada*, *26*, 105–141.

Villarreal, I., & Gil–Sarratea, N. (2020). The effect of collaborative writing in an EFL secondary setting. *Language Teaching Research*, *24*(6), 874–897.

Villarreal, I., & Munarriz-Ibarrola, M. (2021). "Together we do better": The effect of pair and group work on young EFL learners' written texts and attitudes. In M. P. García Mayo (Ed.), *Working collaboratively in second/foreign language learning* (pp. 89–116). De Gruyter.

Vygotsky, L. S. (1986). *Thought and language*. MIT Press.

Watanabe, Y., & Swain, M. (2007). Effects of proficiency differences and patterns of pair interaction on second language learning: Collaborative dialogue between adult ESL learners. *Language Teaching Research*, *11*(2), 121–142.

Wigglesworth, G., & Storch, N. (2009). Pair versus individual writing: Effects of fluency complexity and accuracy. *Language Testing*, *26*, 445–466.
Wolfe-Quintero, K., Inagaki, S., & Kim, H. (1998). Second language development in writing: Measures of fluency accuracy and complexity. University of Hawaii at Manoa.
Woo, M. M., Chu, S. K. W., & Li, X. (2013). Peer-feedback and revision process in a wiki mediated collaborative writing. *Educational Technology Research and Development*, *61*(2), 279–309.
Yang, W., Lu, X., & Weigle, S. C. (2015). Different topics, different discourse: Relationships among writing topic, measures of syntactic complexity, and judgments of writing quality. *Journal of Second Language Writing*, *28*, 53–67.
Zhang, M., & Plonsky, L. (2020). Collaborative writing in face-to-face settings: A substantive and methodological review. *Journal of Second Language Writing*, *49*, 100753.
Zhao, H. H. (2014). Investigating teacher-supported peer assessment for EFL writing. *ELT Journal*, *68*, 155–168.

How to cite this article: Bueno-Alastuey, M. C., Vasseur, R., & Elola, I. (2022). Effects of collaborative writing and peer feedback on Spanish as a foreign language writing performance. *Foreign Language Annals*, 1–23. https://doi.org/10.1111/flan.12611

APPENDIX I: SIX-POINT ANALYTIC SCALE FOR OVERALL QUALITY

Organization and content	**Excellent**	**Good**		**Fair**		**Poor**
Content						
Use of supporting details	6	5	4	3	2	1
Structure and organization						
Contains all the content required in the instructions: intro, body, and conclusion	6	5	4	3	2	1
Information is organized in a logical sequence						
Quality of language	**Excellent**	**Good**		**Fair**		**Poor**
Vocabulary: Uses accurate and appropriate vocabulary Avoids false cognates Spells and accentuates words correctly Attempts to use vocabulary specific to the topic	6	5	4	2		1
Grammar: Correctly conjugates verbs (present, irregulars, gustar, etc.) Uses correct agreement in articles, nouns, and adjectives Uses Spanish sentence structure & avoids poor direct English translation Attempts to use features of language recently learned	6	5	4	2		1
						TOTAL.../24 =____

APPENDIX II: RUBRIC SCORE BREAKDOWN BY CATEGORY

	Group 1			Group 2		
	Pretest	Postcollaborative writing	Postpeer feedback	Pretest	Postpeer feedback	Postcollaborative writing
Content	4.11 (1.1)	4.39 (1.1)	2.61 (1)	3.68 (1.4)	4.59 (1)	4.36 (1.3)
Organization	4.11 (1.3)	3.17 (1.2)	3.56 (1.3)	3.55 (1.5)	3.82 (1.1)	3.32 (1)
Vocabulary	3.28 (1.3)	3.61 (1.3)	4.22 (0.9)	2.82 (1.4)	4.09 (1.2)	3.91 (1.4)
Grammar	3.22 (1.2)	3 (1.3)	3.72 (1.4)	3.14 (1.4)	3.77 (1.5)	4.05 (1.2)
Total	**14.72 (4.2)**	**14.17 (4.3)**	**16.11 (3.7)**	**13.18 (5.1)**	**16.27 (3.9)**	**15.64 (4)**

APPENDIX III: ASSIGNMENT #1: CELEBRANDO LA VICTORIA

Vas a escribir una narrativa (entre 250 o 300 palabras) sobre alguna fiesta o celebración a la que hayas asistido en la que se celebraba el triunfo de un equipo en algún deporte. Recuerda escribir tres párrafos (primero presenta el contexto en el que se llevó a cabo la fiesta, después cuenta que pasó y finalmente cómo terminó). Incluye información sobre:

- dónde tuvo lugar la celebración y qué estaban celebrando
- como fuiste a la fiesta y con quien
- personas que asistieron y ropa que llevaron a la fiesta
- qué hicieron en la fiesta
- comida y bebida que había
- qué pasó en la fiesta
- alguna anécdota divertida sobre la fiesta
- cómo terminó la fiesta y qué pasó después.

Vocabulario útil: Conectores temporales:
El domingo a las...
Después/luego/más tarde...
Finalmente,
De pronto .....
En cuanto .....