# Effects of proficiency and collaborative work on child EFL individual dictogloss writing

Asier Calzada and María del Pilar García Mayo
Universidad del País Vasco | Euskal Herriko Unibertsitatea (UPV/EHU)

Collaborative writing has been traditionally studied in terms of language-related episodes (LREs), which have been shown to be influenced by learner proficiency. Yet, the impact of collaboration on the written product has received less attention, especially regarding child EFL learners. Our study analyzes the individual reconstructions produced by 30 Spanish-Basque EFL children (aged 11–12) before and after (T1 and T3) they completed a collaborative dictogloss (T2). From the analysis of their LREs at T2, we predicted that certain areas (grammar and mechanics) could reflect more changes at T3 than others. Moreover, we wanted to determine whether those changes were moderated by the learners' and their partners' proficiency at T2: low (LP) or high (HP). Text-based and rubric measurements showed that only grammatical complexity improved in children's individual writing from T1 to T3. Regarding proficiency, LP children performed significantly worse than their HP counterparts at T1 and T3 in most writing dimensions. Partner proficiency only influenced accuracy, and unexpectedly, working with an LP partner did not appear to have a detrimental effect. Our findings stress the need to carry out longitudinal studies to further determine the role of collaboration in L2 writing and knowledge development.

**Keywords:** EFL, young learners, dictogloss, L2 writing, collaboration, proficiency

## 1. Introduction

Writing is a particularly challenging task because, as summarized in Yasuda (2019), different skills need to be developed: transcription (spelling and letter formation), language-based skills (word choice, lexical variation/sophistication, construction of grammatically correct sentences, among others) and mechanics

https://doi.org/10.1075/ltyl.20003cal
*Language Teaching for Young Learners* 3:2 (2021), pp. 246–274. ISSN 2589-2053 | E-ISSN 2589-207X

(punctuation). Moreover, writers need to be aware of the type of audience their text is addressed to and of the use of coherence and cohesive devices. This task is even more challenging for second and foreign language (L2) learners (Manchón & Matsuda, 2016), as the process and product are influenced by learner proficiency level in the L2, their literacy in the first language and potential differences in rhetorical approaches to the text.

Although L2 writing research is on the increase, examples of studies on writing in foreign languages continue to be scarce in the literature (Reichelt, Lefkowitz, Rinnert, & Schultz, 2012), and even less so when it comes to school contexts (Lee, 2016; Ortega, 2009). Against this backdrop, we identify two main reasons to advance the research agenda in this area. Firstly, English as a Foreign Language (EFL) settings differ from English as a second Language (ESL) settings (larger class sizes, less exposure to the L2, instructors' greater concern for grammatical accuracy over content) (Reichelt et al., 2012), and secondly, they have been reported to suffer from a lack of systematicity in writing instruction (Matsuda & DePew, 2002).

In this paper we investigate how young EFL learners can be helped to *write to learn* and *learn to write* (Manchón, 2011) through a collaborative writing task, namely, a dictogloss task (Wajnryb, 1990). Previous dictogloss studies have employed a "task-based performance" approach (Plonsky & Kim, 2016), that is, researchers analyzed what kind of interaction resulted from dictogloss, in the form of language-related episodes (LREs) (Swain & Lapkin, 1998) and how individual variables such as proficiency (Leeser, 2004) impacted their discussions. However, in the present study, we will approach dictogloss from a "task-as-treatment" perspective, as our aim will be to examine the impact of those form-focused discussions on individual L2 writing.

To the best of our knowledge, there is no study with young learners (YLs) that has considered how learners' and their partner's proficiency level during a collaborative dictogloss (treatment) impacts on the quality of the individually written dictogloss (pretest and posttest). The current study aims, therefore, to address this gap by analyzing the individual writing of 30 Spanish EFL primary school children (aged 11–12) using a range of text-based and rubric measures.

## 2. Literature review

### 2.1 The impact of collaboration on L2 writing

Storch (2019) defines collaborative writing as "an activity that requires the co-authors to be involved in all stages of the writing process, sharing the respon-

sibility for and the ownership of the entire text produced" (p. 40). Underpinned by Socio-cultural theory (SCT) (Vygotsky, 1978), collaborative writing tasks have been considered a site for knowledge co-construction when learning a language. Learners engage in discussion over language *(languaging*, Swain, 2006), which can help them gain or consolidate L2 knowledge (Storch, 2016). Writing collaboratively can also trigger a pooling of knowledge about language, which Donato (1988) termed 'collective scaffolding'.

A number of studies with adults in the Australian ESL setting investigated the impact of individual vs. collaborative work on the written text. Storch (1999), one of the earliest comparison studies, focused on grammatical accuracy in a small-scale research project with eleven intermediate university students who completed three tasks (a cloze exercise, a composition and a text reconstruction) individually and in pairs on two separate days. Storch reported a positive effect of collaboration on grammatical accuracy. In a subsequent study, Storch (2005) compared intermediate ESL learners' pair ($n=9$ dyads) and individual ($n=5$) work in a writing class. The findings showed that pairs needed more time for the task and produced shorter texts, which were however more accurate and complex than those written by individuals on their own. Examining the oral interactions of pairs, Storch reported that, unlike individual learners, students working in pairs had opportunities to pool their knowledge and provide feedback to one another. In Storch and Wigglesworth (2007) the database was larger (24 pairs and 24 individuals) and the findings showed that there were no differences in terms of the fluency and complexity of the texts but there were statistically significant differences in terms of grammatical accuracy.

Research on the impact of collaboration on written language output has also been carried out in foreign language settings with adult learners. For example, Malmqvist (2005) examined the texts produced by 10 Swedish and 2 Finnish learners of German as a third language, who had English as an L2. She used three dictogloss tasks, the first and the third completed individually and the second in small groups of three. The findings of the study demonstrated that the group discussions the learners held when working collaboratively in small groups did affect their written language output. The collaborative texts were longer, more detailed and syntactically more complex than the ones reconstructed individually. Fernández Dobao (2012) was the first study to compare group, pair and individual work in collaborative writing tasks in the L2 classroom. The study was conducted with six intermediate classes of Spanish as a foreign language (SFL) in the US. Twenty-one learners worked individually, thirty in pairs and sixty in groups of four on a jigsaw task. They had to rearrange the pictures provided and produce a written text. Fernández Dobao examined whether the number of participants had an effect on the fluency, complexity and accuracy of the written products and

on the frequency and nature of the oral interaction produced in pairs and groups. Regarding the former, she reported that the texts written by groups were not only more accurate than those written individually, but they were also more accurate than those written by pairs.

Little research on collaborative work has been carried out in school settings in general and with young learners in EFL contexts in particular. Basterrechea and García Mayo (2013) investigated the effects of collaborative work on the production of the present tense marker *-s* by 41 Content and Language Integrated Learning (CLIL) and 40 EFL learners (age range 15–16, L1 Spanish) during dictogloss. Collaborative text reconstruction led to more accurate use of the target form than individual text reconstruction in the two educational contexts, and CLIL dyads who collaborated outperformed those who worked individually in the same setting.

In a recent study with adolescent EFL learners (age range 16–17, intermediate proficiency level), Villarreal and Gil-Sarratea (2019) explored the learning affordances of collaborative work. In their study, a control group ($n=16$) produced an argumentative text individually and an experimental group ($n=16$) did so in pairs while recording their interactions. In line with previous research, the findings revealed that the pairs produced shorter but more accurate and slightly more lexically and grammatically complex texts. Moreover, texts were analyzed qualitatively using overall quality measures and the findings showed that the pairs also obtained higher scores in content, structure and organization.

In sum, most studies that have used collaborative text reconstruction tasks indicate that collaboration impacts target form production and helps L2 writing by drawing the learner's attention to formal aspects of the language and discourse. The few studies that have compared actual pair vs individual written production have shown that collaboration increases the accuracy of the texts produced as a result of learner interaction during collaboration. However, no research has addressed this important role of collaboration with YLs in EFL primary school settings.

## 2.2 The role of proficiency in collaborative L2 pair work

One of the variables that impacts collaborative pair interaction is proficiency differences between the members of the pair. Leeser (2004) examined pair interaction among 21 dyads of SFL learners completing a dictogloss task targeting aspect distinction in the past tense. The learners were divided into high-high (HH), high-low (HL) and low-low (LL) proficiency pairs. The analysis of their interaction showed that low proficiency learners benefited from being paired with high proficiency learners, but HH pairing seemed to be the optimal setting for focusing on form.

In an ESL setting, Watanabe and Swain (2007) studied the interaction of 12 Japanese learners, in which four core participants interacted with higher and lower proficiency non-core participants in a three-stage writing task. Their findings showed that it was not proficiency but, rather, patterns of pair interaction (Storch, 2002) that influenced production of LREs and post-test performance. That is, when learners worked collaboratively, they were more likely to achieve higher posttest scores regardless of the partner proficiency. Kim and McDonough (2008) support the findings in Leeser (2004) regarding the role of proficiency in pair interaction. In their study, eight Korean L2 learners produced more target-like resolved lexical LREs when they performed collaborative work with an advanced interlocutor instead of with an intermediate interlocutor. Storch and Aldosari (2013) supported both Leeser (2004) and Watanabe and Swain (2007). In their study, 60 Arabic EFL learners, allocated into similar (HH and LL) and mixed (HL) L2 proficiency pairs, completed a short composition. The researchers considered both the learners' overt attention to language use and amount of L2 use as well as their dyadic patterns. Their findings showed that the greater focus on language use occurred among HH pairs, but the authors claimed that patterns of interaction need to be taken into account because HL pairs produced more LREs than LL pairs but only if they formed a collaborative or expert/novice pattern of interaction.

All the studies mentioned above dealt with collaborative dialogue in relation to proficiency. Conversely, Shin, Lidster, Sabraw and Yeager (2016) examined the quality of adult ESL learners' ($n = 38$) joint dictogloss text in terms of content accuracy, operationalized as idea units (Carrell, 1985). Unlike previous research, they used a mixed-method design in which the same students completed equivalent tasks twice but with learners of a different proficiency level. The findings showed that partner proficiency had no significant effect on idea unit gains. However, the general trend was that low proficiency students benefited more from collaboration, especially if paired with higher-level partners (although they also showed the largest variation). The present study will try to fill the gap regarding the impact of proficiency on subsequent individual production in the underexplored young EFL population.

### 2.3 Quantitative and qualitative measures of L2 writing

How to describe learners' performance in an L2 has been a key issue in second language acquisition research (SLA) from the inception of the field (see Larsen-Freeman, 1978). When assessing writing, research has heavily relied on complexity, accuracy and fluency (CAF) measures (Housen, Kuiken, & Vedder, 2012). Statistically they have been shown to be independent constructs and they have been

argued to follow a consistent acquisitional sequence, yet sometimes influenced by learners' personal goals and learning style (Ellis, Skehan, Li, Shintani, & Lambert, 2020). Nevertheless, in the present study only complexity and accuracy will be examined.[1]

To gauge complexity, a problematic construct because of the many ways in which it has been operationalized (Pallotti, 2015), both grammatical and lexical complexity have been considered. Regarding the former, three subdimensions have been identified (Michel, 2017): length (short vs long units, e.g. number of words per clause), variation (variety of units, e.g. number of morphemes used) and interdependence (relations between units, e.g. amount of coordination vs subordination). On the other hand, lexical complexity has been gauged by using type-token based measures, such as Guiraud's (1960) index and D (Meara & Miralpeix, 2017).

Regarding accuracy, how appropriate lexical, grammatical, semantic and pragmatic choices are with respect to the L2 target forms, research with EFL learners has opted for amount of errors per 100 words and grammar errors per 100 words (Tejada-Sánchez & Pérez Vidal, 2018), due to the short compositions low proficiency learners write and the need to tease apart grammar errors from other frequent errors in these learners' written production, such as spelling errors.

Furthermore, when assessing task-based performance, it has been suggested that measures capturing the extent to which learners meet the task goals should also be included (Michel, 2017; Pallotti, 2009). In the present study, learners had to keep the gist of the original dictogloss text in their writing, and therefore, as in Shin et al. (2016), we opted to assess the content accuracy (CA) of learners' reconstructions by quantifying the amount of Idea Units (IUs) retrieved.

The ecological validity of some of these analytic measures has been questioned (McDonough & García Fuentes, 2015) and, therefore, it seems sensible to use overall quality measures as well when assessing L2 written output from instructed SLA. As reported by Polio and Shea (2014), holistic measures are ecologically valid and practical – they are commonly used by EFL primary and high-school teachers. The present study will hence use both quantitative and qualitative measures to examine YLs' written production.

In summary, most studies on collaborative writing have so far primarily focused on the frequency, nature and outcome of LREs in ESL/EFL adult learners' output in relation to proficiency. Yet, very few have analyzed how collaborative dialogue influences L2 writing. Moreover, to the best of our knowledge, this research path is still uncharted regarding the EFL child population.

---

**1.** The exact time spent on the writing task at T1 and T3 was not recorded, and therefore, a fluency measure (e.g. number of words written in a given period of time) could not be obtained.

## 3. The study

### 3.1 Research questions

In the present study, we entertained the following questions:

1. Does young EFL learners' individual written production improve from Time 1 (T1) to Time 3 (T3) after completing a collaborative dictogloss task at Time 2 (T2)?
2. Does learners' proficiency or partner proficiency (high or low) impact on their individual written production (from T1 to T3)?

In order to establish our hypotheses (see below), we looked at pair interaction at T2 (collaborative dictogloss). Our unit of analysis were LREs (Swain & Lapkin, 1998), that is, parts of the learners' interactional conversation where they "talk about the language they are producing, question their language use, or correct themselves or others" (p. 328). As in previous literature about LREs and YLs (Collins & White, 2019; García Mayo & Imaz Agirre, 2019), we classified them according to their linguistic focus:[2] target form (3rd *-s*), other grammatical forms, lexis and mechanics (spelling and punctuation). Apart from tallying the LREs, we calculated their number of turns, in order to have a better estimation of the quality of that talk. In fact, LREs involving more turns usually imply more dyadic engagement and influence learning (Fernández Dobao, 2016). The results were classified according to the learners' grouping proficiency distribution: low-low (LL) ($n = 10$), high-low (HL) ($n = 10$) and high-high (HH) ($n = 10$). Table 1 features this information:

**Table 1.** LREs by focus and proficiency grouping

| | 3rd *-s* | | Other grammar | | Lexis | | Mechanics | |
|---|---|---|---|---|---|---|---|---|
| | Turns | LREs | Turns | LREs | Turns | LREs | Turns | LREs |
| | M (SD) | M (SD) | M (SD) | M (SD) | M (SD) | M (SD) | M (SD) | M (SD) |
| LL ($n = 10$) | 0 | 0 | 2.60 (2.41) | 1 (1) | 3.20 (3.11) | 0.80 (0.84) | 2.20 (1.92) | 1.20 (1.09) |
| HL ($n = 10$) | 0 | 0 | 19.40 (16.30) | 4.20 (2.50) | 8 (2.34) | 1.80 (0.45) | 7 (11.31) | 2.60 (4.77) |
| HH ($n = 10$) | 4.20 (4.55) | 0.80 (0.80) | 18.40 (19.88) | 5 (4.85) | 17.40 (18.24) | 4 (2.55) | 19.40 (14.10) | 7.60 (5.13) |

2. LRE classifications do not include textual level foci (adequacy, coherence and cohesion). We acknowledge that differences shown in any of those stylistic dimensions from T1 and T3 are difficult to relate exclusively to the learners' interaction at T2.

On average, HH generated the most and lengthiest LREs in all focus categories, with the exception of turns about other grammar features, where HL produced more. HH were followed by HL and LL. Yet, the standard deviation values across the three grouping conditions indicate wide differences within each of the groups. In order to check whether inter-group differences were significant, we conducted a one-way ANOVA. The omnibus ANOVA (excluding from the analysis the target form, since there were no values in two of the three grouping conditions) showed that there were only significant differences (with a large effect size) in Lexis LREs ($F(1,12) = 5.43$, $p = .021$; $\eta = .47$). The Tukey post-hoc test showed that there was a significant difference between the HH group and the LL ($p = 0.019$; $d = 1.69$; M difference CI = [0.55, 5.85]), with a large effect of proficiency grouping.

After observing the learners' LREs, we can hypothesize the following for our research questions:

1. There will probably not be an improvement in child learners' individual 3rd person *-s* accuracy in written production, as they did not focus their attention so much on this target form. On the other hand, there might be some improvement in the rest of the linguistic domains, especially in other grammatical forms or mechanics, as, on average, they produced more LREs.
2. The comparison of the mean results from HH and LL suggests that proficiency played a role, as the former generated a higher amount of discussions than the latter. Therefore, we expect that the writing gains will be greater for high proficiency learners than for low proficiency ones, especially in the domain of lexis. With regards to partner proficiency, the heterogeneous condition's (HL's) mean results also suggest there might be some sort of effect of this variable. Low proficiency learners might have benefitted from the larger amount of LREs held with their higher proficiency peers. Therefore, they might make more gains in their writing scores at T3 than the low proficiency learners who had other low proficiency learners as peers and who generated fewer LREs. Likewise, we could also foresee that high proficiency learners who worked with low proficiency learners may have not benefitted from T2 as much as those working with other high proficiency learners.

### 3.2 Design and procedure

As part of a larger study, written data were collected three times throughout three consecutive weeks from 67 Spanish-Basque EFL learners aged 11–12. They were all from the same school and belonged to three parallel classes of 6th year of Primary education. Although Spanish was the dominant language outside school, the

children were enrolled in a Basque immersion model, also known as "D model" (i.e. all subjects except for English and Spanish were delivered through this language) (Etxeberria & Etxeberria, 2015). Additionally, following a CLIL approach (Dalton Puffer, 2011), a few subjects (1.5h Science and 2h Arts and Crafts) were also delivered through English. Together with the mainstream EFL classes (3h), English exposure added up to 7h per week.

The writing tasks were completed in their school premises during mainstream English lessons. Learners from the three parallel classes were never mixed for the experimental procedure. For the current study, we present a subset of that data belonging to thirty learners ($n=30$). In week 1 and week 3, referred to as Time 1 (T1) and Time 3 (T3), learners completed an individual dictogloss task (explained below), whereas at Time 2 (T2) they carried out the collaborative dictogloss in researcher-selected pairs. The main source of data, hence, comes from the individual written reconstructions at T1 and T3 ($n=60$), as these stages served as a pretest and posttest to determine the impact of a collaborative stage in between. The written output at T2 is excluded from the analysis of this study.

*Proficiency test*

All 67 children participating in the larger dictogloss study took, prior to T1, a *Flyers* test (Grammar, Vocabulary and Listening papers) (Cambridge Assessment English, 2018) to assess their English proficiency level. The results of the test indicated that the data were symmetrical and not affected by outliers ($M=73.65$, Mdn = 75.20). Based on the Cambridge Assessment criteria, the *Flyers* raw test scores were translated into shields (from 1 to 5). Consequently, 80% of the scores or more translated into 5 shields, implying that those children had reached the A2 Common European Framework of Reference (CEFR) proficiency level and were ready to move on to the next stage. Conversely, those who had obtained 4 shields (60–80% of the scores) had still room for improvement, and those in the 1–3 shield range were closer to an A1 and pre-A1 level. Therefore, in the present study, those who had scored 80% or more were considered high proficiency students (HP), whilst those below that benchmark were considered the low proficiency ones (LP).

Apart from analyzing the difference in the written output between HP and LP at T1 and T3, as in Shin et al. (2016), we wanted to determine the impact of the partner proficiency variable, that is, whether working with a high or low proficiency partner (HPP or LPP) at T2 (collaborative dictogloss) could have any influence on the written outcome. We controlled for the proficiency distribution within the pairs, in order to avoid excessive differences within homogeneous proficiency pairs (10-point maximum difference between their results in the *Flyers*) and to make sure the difference was wide enough in the heterogeneous setting

(10-point minimum difference). Finally, only 15 pairs succeeded in meeting these requirements, which represent the final sample of the current study.

In total, regarding the first factor (proficiency) there were 15 HP and 15 LP learners, and regarding the second factor (partner proficiency) there were 15 HPP and 15 LPP. In order to ascertain the equivalence between the two levels of both factors before the experimental procedure took place, we followed Larson-Hall's (2016) recommendation outlined by Tryon (2001). We calculated the descriptive statistics and the inferential confidence intervals (Infer CIs) resulting from the *t*-test. Table 2 summarizes the results:

**Table 2.** Flyers test scores and inferential confidence intervals by proficiency and by partner proficiency

| | M (SD) | Infer CI |
|---|---|---|
| LP ($n=15$) | 61.65 (10.51) | [45.80, 77.50] |
| HP ($n=15$) | 87.92 (15.86) | [71.01, 104.83] |
| LPP ($n=15$) | 71.23 (15.38) | [67.74, 74.71] |
| HPP ($n=15$) | 78.35 (16.05) | [74.71, 81.98] |

In the Infer CIs for proficiency we can see that the values overlap, and hence the groups were statistically the same prior to the experimental stages. However, this is not the case for the second factor (partner proficiency), so we can consider the two groups statistically different.

### *The writing task*

In this study, learners had to carry out three dictogloss tasks (Wajnryb, 1990), where the main goal was to reconstruct a text they heard by keeping the gist of the story. Dictogloss has been claimed to be effective for focusing attention on formal aspects of language in the case of adult learners (Alegría de la Colina & García Mayo, 2007), teenagers (Basterrechea & Leeser, 2019; Swain & Lapkin, 2001) and, only recently, with the underexplored population of young EFL learners (Calzada & García Mayo, 2020a; 2021). Furthermore, collaborative dictogloss has been shown to be an enjoyable task for children, as they tend to feel less and anxious and supported by their peers when working with their peers (Calzada & García Mayo, 2020b).

In our tasks, learners had to listen twice to a short recording of a narrative text. During the second listening, they were encouraged to write down some of the key ideas of the text, so that they could resort to them afterwards in the reconstruction stage. Dictogloss is classified as a focused task by Storch (2016). In our

case, it was the third person singular *-s* marker we wanted learners to focus on. This morpheme's low semantic load and phonological salience, as well as morphosyntactic redundancy, make it especially difficult to grasp for ESL and EFL learners (Basterrechea & Leeser, 2019).

Three texts were created "ad-hoc" for the larger experimental procedure (Calzada & García Mayo, 2021) and are available in the IRIS database (Marsden, Mackey & Plonsky, 2016). The main reason to administer three different texts was to avoid same task repetition negative effects (feeling of repetitiveness and boredom), and, what is more, procedural task repetition has been shown to have some positive impact on the oral CAF in young EFL learners (Lázaro-Ibarrola & Hidalgo, 2017; Sample & Michel, 2015). The English teacher informed us that the text genre which learners were most familiar with was narration. This is also the type of text learners of this age are usually required to produce in the diagnostic tests administered by the regional government. Hence, we decided that all texts should take the form of a short story as much as possible. Regarding text characteristics, each of them consisted of 122 words and contained 15 instances of the target 3rd person singular *-s*. Table 3 includes some more text characteristics which ensured their resemblance:

**Table 3.** Dictogloss text characteristics

| | Flesch–Kincaid readability test | Guiraud | Recording time | Recording pace (words/minute) |
|---|---|---|---|---|
| Sweet Surprise | 3.7 | 7.06 | 01:12 | 101.67 |
| Naughty Laura | 3.6 | 6.52 | 01:17 | 95.06 |
| Halloween Night | 2 | 6.85 | 01:12 | 102.50 |

As can be seen, the features remained practically the same across the three texts. The topics were chosen according to the English syllabus for that semester, so that the child learners could be as familiar as possible with them. They dealt with cooking and celebrations. Finally, the recordings were done by the same English L2 speaker.

In order to reduce any potential effect produced by the texts on the written production, the texts were presented in a latin-square design, that is, in each of the three 6th-year primary classes there was a different dictogloss text at a time. Learners could use up to 15–20 minutes to reconstruct the text at T1 and T3, and up to 25 at T2 (collaborative). Besides, they could not consult any external

resources (dictionaries, asking the teacher... etc.) during the task. At T1 and T3 learners carried out the task in their classrooms simultaneously, whereas at T2 the first author and other research assistants took one pair at a time to a separate room to perform the collaborative task and record their interaction with better sound quality standards.

### 3.3 Data coding and analysis

All 60 texts from T1 and T3 were transcribed on Word and on CLAN (MacWhinney, 2000). Guided by previous research on EFL learner text analysis, we employed the following text-based and analytic rubric measures.

*Analytic measures*

*Complexity*

- Grammatical complexity (W/C): considering the clause length (Michel, 2017), we used words per clauses (W/C), that is, the number of words learners produced divided by the total number of clauses in each writing. Based on a linguistic definition, a clause was understood "as a unit consisting of a subject (explicit or implied) plus a predicate, i.e. construction with a finite or non-finite predicator or verb as its nucleus" (Bulté & Housen, 2014, p. 48). Therefore, clauses can be both independent and dependent. Example (1) illustrates clause boundaries (/) in a learner's text (L2 errors have been kept as in the original):

(1) *Every year when is halloween night / goes with his mask wich to houses. / When she goes / said truck or trick / and the people of the houses throw the sweets. / Lucy takes the sweets / and eatit with his sister Ana. / Lucy's mouth it was hill / and goes with his mum to the dentist. / The dentist said / that it was very hill. / Lucy now doesn't eat much sweets.* [S26, HP, HPP, T1]

- We used the frequency options in CLAN to obtain the number of clauses and tokens (words), and we calculated the final W/C measure on Excel.
- Lexical complexity (Guiraud): we used the Guiraud Index for lexical diversity, which is calculated by dividing the number of types by the square root of tokens (obtained from CLAN). According to Meara and Miralpeix (2017), it provides a better estimate of lexical diversity than the traditional type/token ratio, as it limits to a certain extent the text size effect, and can be used with short texts, as opposed to D. We counted as a type any word which orthographically resembled the original word enough to convey its lexical meaning.

*Accuracy*

- 3rd *-s* marker accuracy rate (3S): being our target feature, we calculated the number of correctly produced instances of this form divided by the number of obligatory contexts in each writing. We used the same software procedure as in W/C to obtain this measure.
- Other grammatical errors per 100 words (GramErr100): apart from the target form, we calculated how many grammatical errors the learners produced in their writing in relation to the number of words. Children in this study produced a wide variety of errors (subject and object dropping, irregular past verb forms, articles, etc.). Once again, we followed the aforementioned software procedure.

*Content accuracy*

- Idea Units (IUs): to assess how successfully our young learners fulfilled the main requirement of the dictogloss task (getting the gist of a story), both authors divided the original texts into ten idea units (Carrel, 1985). Then, we assessed how many of those original ideas each of the learner's text retained. Although the total number of ideas gathered in the text may rely on linguistic competences other than writing (such as listening comprehension and vocabulary knowledge) we considered that at T2 learners could share some cognitive strategies to recall as many IUs as possible which would help them in the subsequent individual writing stage (T3).

*Overall text quality rubric measures*

We used a rubric (see Appendix) that had been previously used in the literature to assess child EFL learners' texts (Villarreal & Munarriz-Ibarrola, 2021). It was partly based on the rubric used in the regional government's diagnostic tests for determining young learners' competence in English writing (Department of Education, 2020), and it had been occasionally employed by the English teacher at the school for assessment. It consists of six dimensions: adequacy (Adq), coherence (Coher), cohesion (Cohes), grammatical accuracy (Acc), mechanics (Mech), and lexical range (Lex), rated from 1 (lowest) to 3 (highest). The scale includes some descriptors to help the raters make their choices.

*Interrater reliability*

For the most subjective measures (overall quality ratings and IUs), the whole set of writings ($n=60$) was evaluated by both authors. Table 4 summarizes the

descriptive statistics and the Krippendorff's alpha[3] for each measurement at T1 and T3:

**Table 4.** Interrater descriptive statistics and Krippendorff's alpha for the subjective measures

| | T1 ($n=30$) | | | T3 ($n=30$) | | |
|---|---|---|---|---|---|---|
| | **Rater 1 M (SD)** | **Rater 2 M (SD)** | **Krippendorff's alpha** | **Rater 1 M (SD)** | **Rater 2 M (SD)** | **Krippendorff's alpha** |
| Adequacy | 1.87 (0.74) | 1.63 (0.72) | 0.75 | 1.85 (0.63) | 1.53 (0.82) | 0.53 |
| Coherence | 2.02 (0.67) | 1.70 (0.69) | 0.67 | 1.90 (0.56) | 1.60 (0.72) | 0.52 |
| Cohesion | 1.70 (0.62) | 1.40 (0.62) | **0.45** | 1.75 (0.50) | 1.33 (0.55) | **0.27** |
| Gram. accuracy | 1.75 (0.69) | 1.53 (0.63) | 0.71 | 1.50 (0.56) | 1.20 (0.48) | **0.31** |
| Mechanics | 1.93 (0.69) | 1.72 (0.69) | 0.64 | 1.97 (0.76) | 1.60 (0.72) | 0.53 |
| Lexical range | 1.90 (0.65) | 1.60 (0.67) | 0.62 | 1.82 (0.64) | 1.47 (0.63) | 0.56 |
| IUs | 4.52 (2.30) | 4.43 (2.55) | 0.58 | 4.53 (2.09) | 4.32 (2.62) | **0.45** |

*Note*: In bold, low reliability coefficients

The interraters' agreement was considered moderate in most of the cases (0.5–0.6) and relatively high in exceptional cases, such as adequacy at T1 (> 0.7). Therefore, the mean rate was calculated to obtain the definite value for each dimension. However, in four cases the agreement was low (cohesion at T1, and cohesion, grammatical accuracy and IUs at T3) (< 0.5). In those four cases, the raters went through all 60 writings together and provided an agreed score.

Regarding objective measures (W/C, Guiraud, 3S, GramErr100), both authors tried coding some writings from the database of the larger study which were not part of the present one. After defining how to code errors, clauses and types, the first author carried out the analysis for the 60 writings from the current study.

**3.** For the rubric dimensions, Krippendorff's alpha for ordinal data was calculated, whereas for Ideas Units we used the same alpha for ratio data. The calculations were done using the online software ReCal (Freelon, 2013).

## 3.4 Results

To examine the interaction between time, proficiency and partner proficiency and the results in the different writing measures, both between and within subjects, we opted for a Repeated-Measures ANOVA. Before conducting the analysis, we checked the assumptions of normality and homogeneity of variance on SPSS 24 (IBM Corp, 2016). Prior to this analysis, the data from the rubric scale was normalized, so that it better suited the statistical test assumptions.

With regards to normality, we checked the distribution at T1 and T3. In both grouping factors (proficiency and partner proficiency) this assumption was generally not met, as indicated by the Shapiro-Wilk test ($p < .05$). However, as previous simulation studies show, ANOVAs can work well with not normally distributed data (Blanca, Alarcón & Arnau, 2017), since they are mostly affected by homogeneity of variance.

The second assumption, the homogeneity of variance, as shown by the results of the Levene's test, was largely met ($p > 0.5$) at both times and in both grouping condition factors. Table 5 and 6 show the descriptive statistics[4] for each grouping condition at T1 and T3, respectively.

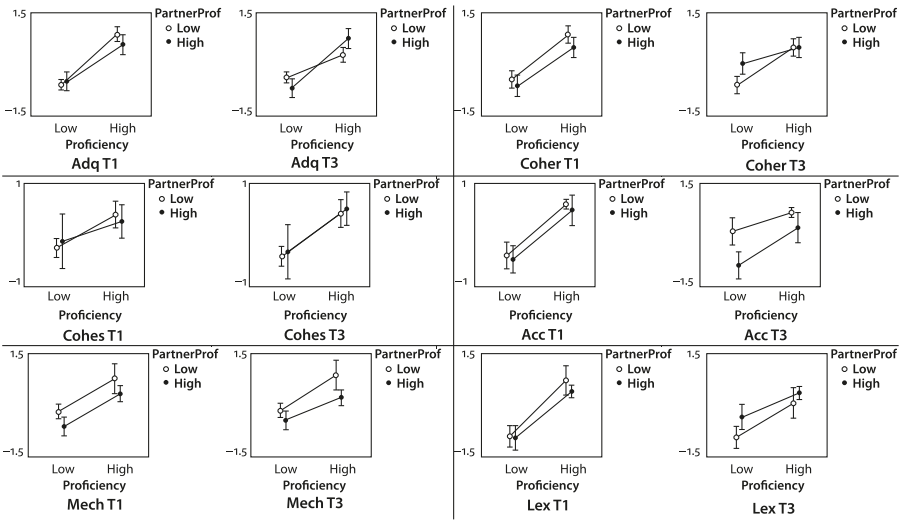


**Figure 1.** Interaction plots of the repeated-measures ANOVA at T1 and T3 for the rubric dimensions in standardized values. Error bars show standard error

4. The descriptive statistics for the rubric results are not standardized for a better interpretation of the results

**Table 5.** Descriptive statistics for T1

| | | Low proficiency | | | High proficiency | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | *Partner proficiency* | | Total | *Partner proficiency* | | Total | Total partner proficiency | |
| | | *Low* (*n* = 10) | *High* (*n* = 5) | LP (*n* = 15) | *Low* (*n* = 5) | *High* (*n* = 10) | HP (*n* = 15) | LPP (*n* = 15) | HPP (*n* = 15) |
| | | M (SD) | M (SD) | M (SD) | M (SD) | M (SD) | M (SD) | M (SD) | M (SD) |
| Rubric | Adq | 1.28 (0.38) | 1.35 (0.65) | 1.30 (0.46) | 2.35 (0.74) | 2.12 (0.53) | 2.20 (0.59) | 1.63 (0.72) | 1.87 (0.67) |
| | Coher | 1.53 (0.51) | 1.40 (0.65) | 1.48 (0.54) | 2.40 (0.76) | 2.15 (0.38) | 2.23 (0.52) | 1.82 (0.72) | 1.90 (0.59) |
| | Cohes | 1.25 (0.26) | 1.30 (0.45) | 1.26 (0.32) | 1.50 (0) | 1.45 (0.50) | 1.47 (0.40) | 1.33 (0.24) | 1.40 (0.47) |
| | Acc | 1.35 (0.46) | 1.30 (0.54) | 1.33 (0.47) | 2 (0.92) | 1.92 (0.47) | 1.95 (0.62) | 1.57 (0.69) | 1.72 (0.56) |
| | Mech | 1.45 (0.60) | 1.20 (0.27) | 1.37 (0.52) | 2 (0.71) | 1.75 (0.35) | 1.83 (0.49) | 1.63 (0.67) | 1.57 (0.42) |
| | Lex | 1.38 (0.41) | 1.35 (0.65) | 1.37 (0.48) | 2.25 (0.61) | 2.07 (0.44) | 2.13 (0.49) | 1.66 (0.63) | 1.83 (0.61) |
| Complexity | W/C | 5.87 (3.11) | 6.96 (5.58) | 6.23 (3.93) | 4.41 (1.55) | 11.89 (5.29) | 9.40 (5.66) | 5.38 (2.72) | 10.24 (5.72) |
| | Guiraud | 4.30 (0.52) | 4.08 (0.47) | 4.23 (0.50) | 5.10 (0.52) | 5.08 (0.60) | 5.08 (0.55) | 4.57 (0.64) | 4.75 (0.73) |
| Accuracy | 3S | 0.15 (0.25) | 0.31 (0.30) | 0.20 (0.27) | 0.44 (0.39) | 0.40 (0.30) | 0.41 (0.32) | 0.24 (0.32) | 0.37 (0.29) |
| | GramErr100 | 38.90 (9.16) | 33.67 (15.36) | 37.16 (11.31) | 22.12 (10.74) | 37.03 (19.02) | 32.06 (17.85) | 33.31 (12.41) | 35.91 (17.40) |
| CA | IUs | 3.25 (1.90) | 2.20 (2.36) | 2.90 (2.04) | 6 (1.87) | 6.07 (1.36) | 6.05 (1.48) | 4.17 (2.26) | 4.78 (2.52) |

The Repeated-measures ANOVA for each of the measurements allowed us to check whether there were within- or between-group differences, as well as any potential interactions between time, proficiency and partner proficiency. The interaction plots for each dependent variable, generated on JASP (JASP Team, 2019), are provided, in Figure 1 and 2. The inferential statistics showed a significant main effect of proficiency in almost all measurements (Adq: $F(1, 26) = 24.27$, $p < .001$, $\eta = .48$; Coher: $F(1, 26) = 17.49$, $p < .001$, $\eta = .40$); Cohes: $F(1, 26) = 6.10$, $p = .020$, $\eta = .19$; Acc: $F(1, 26) = 7.19$, $p = .013$, $\eta = .21$; Mech: $F(1, 26) = 10.51$, $p = .003$,

**Table 6.** Descriptive statistics for T3

| | | Low proficiency | | | High proficiency | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | *Partner proficiency* | | Total | *Partner proficiency* | | Total | Total partner proficiency | |
| | | *Low* (*n* = 10) | *High* (*n* = 5) | LP (*n* = 15) | *Low* (*n* = 5) | *High* (*n* = 10) | HP (*n* = 15) | LPP (*n* = 15) | HPP (*n* = 15) |
| | | M (SD) | M (SD) | M (SD) | M (SD) | M (SD) | M (SD) | M (SD) | M (SD) |
| Rubric | Adequacy | 1.38 (0.36) | 1.15 (0.14) | 1.30 (0.32) | 1.85 (0.78) | 2.20 (0.75) | 2.08 (0.75) | 1.53 (0.56) | 1.85 (0.80) |
| | Coherence | 1.40 (0.46) | 1.80 (0.45) | 1.53 (0.48) | 2.10 (0.55) | 2.10 (0.66) | 2.10 (0.60) | 1.63 (0.58) | 2 (0.60) |
| | Cohesion | 1.35 (0.41) | 1.40 (0.42) | 1.37 (0.40) | 1.80 (0.45) | 1.85 (0.58) | 1.83 (0.52) | 1.50 (0.46) | 1.70 (0.56) |
| | Accuracy | 1.45 (0.60) | 1 (0) | 1.30 (0.53) | 1.70 (0.97) | 1.50 (0.67) | 1.57 (0.75) | 1.53 (0.72) | 1.33 (0.59) |
| | Mechanics | 1.65 (0.47) | 1.50 (0.50) | 1.60 (0.47) | 2.20 (0.57) | 1.85 (0.41) | 1.97 (0.48) | 1.83 (0.56) | 1.73 (0.46) |
| | Lexis | 1.40 (0.46) | 1.70 (0.45) | 1.50 (0.46) | 1.90 (0.65) | 2.05 (0.60) | 2 (0.60) | 1.57 (0.56) | 1.93 (0.56) |
| Complexity | W/C | 9.49 (5.09) | 9.36 (3.58) | 9.44 (4.51) | 9.14 (6.42) | 17.15 (9.66) | 14.48 (9.33) | 9.37 (5.33) | 14.55 (8.84) |
| | Guiraud | 4.46 (0.70) | 4.53 (0.44) | 4.48 (0.61) | 5.14 (0.69) | 5.55 (0.67) | 5.41 (0.68) | 4.68 (0.75) | 5.21 (0.77) |
| Accuracy | 3S | 0.21 (0.29) | 0.08 (0.12) | 0.17 (0.25) | 0.45 (0.44) | 0.23 (0.25) | 0.30 (0.33) | 0.29 (0.35) | 0.18 (0.22) |
| | GramErr100 | 36.54 (12.68) | 48.94 (14.99) | 40.67 (14.29) | 24.84 (13.22) | 39.72 (12.22) | 34.76 (14.09) | 32.64 (13.63) | 42.79 (13.43) |
| CA | IUs | 3.50 (1.51) | 2.50 (1.66) | 3.17 (1.58) | 4.60 (2.50) | 6.30 (1.90) | 5.73 (2.19) | 3.87 (1.88) | 5.03 (2.56) |

$\eta = .27$; Lex: $F(1,26) = 13.89$, $p < .001$, $\eta = .34$; Guiraud: $F(1,26) = 17.05$, $p < .001$, $\eta = .39$; 3S: $F(1,26) = 4.47$, $p < .044$, $\eta = .14$; GramErr100: $F(1,26) = 4.49$, $p = .044$, $\eta = .12$; IUs: $F(1,26) = 22.69$, $p < .001$, $\eta = .44$). Therefore, high and low proficiency learners were significantly different from each other at both testing times in those dimensions. HP learners scored on average higher than LP in all the rubric dimensions at T1 and T3. Besides, they also performed better in lexical complexity, 3S and Idea Units. Conversely, as expected, LP produced significantly more grammar errors than HP regardless of the testing time. The effect size values indi-

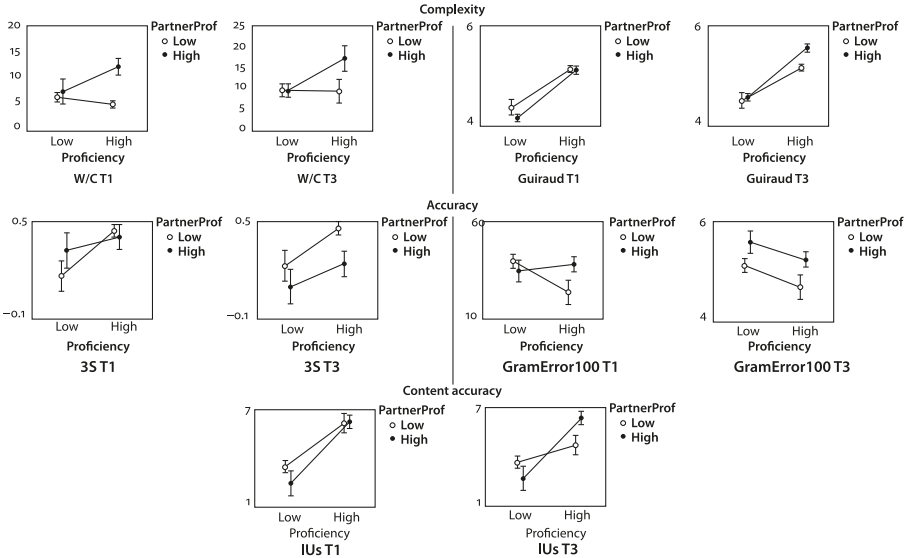


**Figure 2.** Interaction plots of the repeated-measures ANOVA at T1 and T3 for the text-based dimensions. Error bars show standard error

cate that a medium to large proportion of variance in the data is explained by the proficiency factor (small: $\eta = .06$; medium: $\eta = .16$; large: $\eta = .36$).[5] Hence, proficiency could account for as little as 12% of the variance in GramErr100, and as much as 48% in Adequacy.

Our second independent factor, partner proficiency, was only shown to have a significant main effect with a medium effect size in the between-group comparison in GrammError: $F(1, 26) = 5.20$, $p = .031$, $\eta = .14$. This means that learners who worked with a LPP at T2 generated significantly fewer grammatical errors at T1 and T3 than those who worked with a HPP, regardless of their own proficiency.

Time, the within-group factor, only proved to have a significant main effect in W/C, yet with a small effect size: $F(1, 26) = 7.06$, $p = .013$, $\eta = .09$). In other words, children, regardless of their proficiency and partner proficiency, produced significantly more grammatically complex clauses at T3 than at T1. Furthermore, in the case of W/C the proficiency * partner proficiency interaction was also found statistically significant (with a small effect size): $F(1, 26) = 4.60$, $p = .04$, $\eta = .07$). A Tukey post-hoc test was run in order to determine which factor levels were different between each other. Regardless of the testing time, HP

5. These benchmarks are obtained following Plonsky and Oswald's (2014, p. 894) recommended procedure for calculating the typical eta square values in L2 research from *r* value percentiles found in their meta-analytic study (small $r = .25$; medium $r = .40$, large $r = .60$).

learners working with HPP were statistically different from HP working with LPP ($p = .015$, M difference CI = [0.96, 14.52]) and LP working with LPP ($p = .008$, M difference CI = [1.31, 12.38]).

## 4. Discussion and conclusions

We set out to investigate whether a collaborative dictogloss task produced changes in the individual written output of young EFL learners, considering their proficiency and their partner's proficiency at T2 (low or high).

For our first research question, we hypothesized what domains would be improved and which would not from the analysis of the pairs' interaction at T2. We only found a significant change from T1 to T3 in the grammatical complexity measure (W/C). This increase could suggest that the collaborative dialogue generated at T2, especially that related to grammar and mechanics, could have led child learners to shift from short and simple independent clauses to longer compound clauses (by means of both coordination and subordination). In Example (2), the two reconstructions by the participant S3 (LP, LPP) corresponding to T1 and T3 are provided:

(2) T1. Mum prepare a birthday party for Mary / in the morning mum goes to supermarket to buy / mum prepares sweets and cupcakes / at 6 o'clock they eat together the sweets and cupcakes.
*4 independent simple clauses* (31 words)

T3. Every day Lucy prepare hallowen night. / In Hallowen night she put a witch mask / and she ask to trick or treat. / Her mum and she are very scared. / Lucy have wrong the theet / her mother talk to the dentist / and the dentist said / she eat a lot of sweets.
*3 independent simple clauses, 4 independent coordinated clauses, 1 dependent subordinate clause* (49 words)

The examples in (2) illustrate that, while the text at T1 only contains simple clauses, at T3 the participant is able to include compound structures by means of coordination and subordination. Nonetheless, apart from the effect of the collaborative dictogloss, the possible impact of procedural task repetition should not be overlooked either. In fact, child participants carried out the same task type three times (varying from a monologic to a dialogic condition). According to Bygate (2009, 2018), the first time that learners perform a task, they will be more inclined to concentrate on the task meaning and outcome than on form (that is, getting the message across and completing their task requirement). Conversely, the second (or subsequent) time, learners will be able to resort to their experience and

memory of their first performance and, in addition to being more fluent, they will devote more attentional resources to grammar and morphosyntax, hence producing a more complex language. Skehan (2016, 2018) also predicts that repetition allows learners to monitor their production and perform better, although some trade-off effects between linguistic complexity and accuracy are foreseen (Skehan, 2009). Our results tentatively indicate a certain degree of trade-off between these two dimensions. Although no statistical difference was found between accuracy at T1 and T3, the mean rates indicate a downward pattern in HP and LP learners. In contrast, both complexity measures showed the opposite trend (although it did not reach significance in the case of Guiraud).

The fact that collaboration did not have a stronger influence on child learners individual writing could also be related to two other reasons. First, a "one-shot" collaborative dictogloss may not be enough to generate changes in young learners' writing, as has been the case in other studies which looked at changes in their grammar knowledge before and after the completion of the same task (Calzada & García Mayo, 2020a). Secondly, we must also acknowledge that our small sample's negative influence on statistical power could have caused a Type 2 error (Larson-Hall & Herrington, 2010).

Regarding our second research question, our findings allowed us to see an impact of proficiency on the individual dictogloss both at T1 and at T3, but not only in the domain of lexis, as we had initially predicted from the LRE analysis. Moreover, HP learners significantly outperformed LP learners in most dimensions (all rubric dimensions, as well as lexical diversity, grammatical errors and Idea Units). Although Shin et al. (2016) found no effect of partner proficiency in the content adult learners were able to retrieve in their collaborative writing, in our case it was children's own proficiency that played a significant role in retaining IUs from the original texts. As we predicted, since these texts were presented aurally, there may be other competences related to a successful performance in the reconstruction stage, such as listening comprehension or vocabulary knowledge, which are more important than cognitive strategies which we expected to be shared at T2 between high and low proficiency learners.

As far as partner proficiency is concerned, we could not find a significant impact on any of the writing dimensions except for the proportion of grammatical errors per one hundred words (GramError100). We hypothesized that, given the higher number of grammar-related LREs at T2 in the case of HH and HL pairings, learners from these two groupings could possibly benefit from their interaction in subsequent individual writing. Nevertheless, the direction of the impact of partner proficiency was the opposite, as LPP learners produced a lower number of grammatical errors than HPP at T1 and T3. Furthermore, at both testing times, it was the HP children who worked with an LP partner at T2 that obtained the best

grammatical accuracy rates. That is, working with an LP partner did not seem to be detrimental for HP's accuracy, supporting the benefits of expert-novice peer interaction shown for adult L2 learning (Lantolf, 2012); a HP learner can, indeed, detect errors in the LP's production and provide feedback by means of reformulation or recasts (Dao & McDonough, 2017). Last but not least, as previous research has pointed out, proficiency pairing might not be as determinant in peer interaction as patterns of interaction (Storch, 2002; Watanabe & Swain, 2007), which should be studied in future research.

The only significant interaction was found between proficiency and partner proficiency in the W/C (grammatical complexity) variable. HP learners who worked with HPP learners at T2 already produced significantly more complex clauses at T1 than the other pairing conditions, and what is more, they were able to maintain that advantage at T3 despite the general increase in W/C from T1 to T3 discussed above. In other words, collaboration did not homogenize the grammatical complexity of children's writing, and those HP learners, who already produced long clauses at T1, were far from reaching a ceiling. Surprisingly, those HP learners who worked with LP learners at T2 obtained the lowest W/C rates at T1 and T3. Once again, we could explain this difference by suggesting a potential trade-off effect between complexity and accuracy, since this same subgroup of participants scored the highest accuracy rates (3S and GramErr100) at T1 and T3.

However, the fact that these two HP subgroups of learners were prioritizing two different linguistic dimensions (complexity and accuracy) could also respond to individual differences related to children's risk-taking attitude (that is, trying to produce more complex language at the expense of making mistakes). This possibility has been suggested in a previous longitudinal study analyzing YLs' oral CAF, where accuracy was reported to develop while the opposite was true of complexity (Bret Blasco, 2014). Hence, it would be interesting to tap into learners' beliefs about L2 writing to determine the extent to which they are influencing children's linguistic choices.

We should acknowledge some limitations of this study. First and foremost, the lack of a comparison group which completed all three dictogloss tasks individually prevents us from making strong claims about the impact of the collaborative stage. Given that we only found a significant change from T1 to T3 in the case of W/C, it would be interesting to determine whether a comparison group shows the same trend or, instead, indicates some decline or even stronger gains in their writing scores across time. This would certainly shed light on the role of collaboration in subsequent L2 writing. Secondly, our criterium for classifying learners' proficiency dichotomously as high or low, while based on a standardized assessment (Cambridge Assessment English, 2018), may have in some cases amplified children's differences excessively (especially, when their scores were close to the cut-

ting point). Hence, future studies aiming to determine the impact of proficiency on L2 writing at beginner levels may prefer to select only those learners who are clearly either at an A1 or at an A2 level. Last but not least, regarding the overall quality assessment tool, although it was more ecologically valid than the text-based quantitative measures (as it resembled the evaluation instrument used by the school teacher), the interrater reliability coefficients did not report high levels of agreement. Thus, we could question the validity of the rubric for assessing EFL writing in response to a dictogloss task (i.e. the descriptors could lead to confusion or be too vague).

To conclude, the present study has pointed out that assessing YLs' written production can be a rather complex task, due to the high number of dimensions involved in the analysis and the fact that researchers are dealing with a dynamic competence at this learning stage. It has also determined a clear impact of proficiency on individual writing. Our results also suggest that in order to ascertain the impact of collaboration on individual writing from a "task-as-treatment" approach, a single task of this kind is probably not enough to generate significant changes in children's language knowledge and writing expertise. Finally, while partner proficiency failed to show a significant impact, it was interesting to note that in some cases the expected hypotheses and the obtained results were opposite to each other. Further research on this topic is needed adopting a longitudinal approach.

## Funding

This work was supported by the Spanish Ministry of Economy and Competitiveness (MINECO) and National Research Agency and European Regional Development Fund (AEI/FEDER/EU) under Grant FFI2016-74950-P, and by the Basque Government under grant IT904-16.

## Acknowledgements

The authors are grateful to the anonymous reviewers for their insightful comments. Moreover, we would like to thank the school that allowed access to the students and, of course, the students themselves as, without their participation, the study would have not been possible. Thanks also go to Alys Williams for proofreading the manuscript.

## References

Alegría de la Colina, A., & García Mayo, M. P. (2007). Attention to form across collaborative tasks by low-proficiency learners in an EFL setting. In M. P. García Mayo (Ed.), *Investigating tasks in formal language learning* (pp. 91–116). Clevedon: Multilingual Matters.

Basterrechea, M., & García Mayo, M. P. (2013). Language-related episodes during collaborative tasks: A comparison of CLIL and EFL learners. In K. McDonough & A. Mackey (Eds.), *Second language interaction in diverse educational contexts* (pp. 25–43). Amsterdam: John Benjamins. https://doi.org/10.1075/lllt.34.05ch2

Basterrechea, M., & Leeser, M. J. (2019). Language-related episodes and learner proficiency during collaborative dialogue in CLIL. *Language Awareness*, 1–19. https://doi.org/10.1080/09658416.2019.1606229

Blanca, M. J., Alarcón, R., & Arnau, J. (2017). Non-normal data: Is ANOVA still a valid option? *Psicothema*, *29*(4), 552–557. https://doi.org/10.7334/psicothema2016.383

Bret Blasco, A. (2014). L2 English young learners' oral production skills in CLIL and EFL settings: A longitudinal study (Unpublished doctoral dissertation). Universitat Autònoma de Barcelona, Spain.

Bulté, B., & Housen, A. (2014). Conceptualizing and measuring short-term changes in L2 writing complexity. *Journal of Second Language Writing*, *26*, 42–65. https://doi.org/10.1016/j.jslw.2014.09.005

Bygate, M. (2009). Effects of task repetition on the structure and control of oral language. In K. Van den Branden, M. Bygate, & J. M. Norris (Eds.), *Task-based language teaching* (pp. 249–274). Amsterdam: John Benjamins. https://doi.org/10.1075/tblt.1.15eff

Bygate, M. (Ed.). (2018). *Learning language through task repetition*. Amsterdam: John Benjamins. https://doi.org/10.1075/tblt.11

Calzada, A., & García Mayo, M. P. (2020a). Child EFL grammar learning through a collaborative writing task. In W. Suzuki & N. Storch (Eds.), *Languaging in language learning and teaching: A collection of empirical studies* (pp. 20–39). Amsterdam: John Benjamins. https://doi.org/10.1075/lllt.55.01cal

Calzada, A., & García Mayo, M. P. (2021). Child learners' reflections about EFL grammar in a collaborative writing task: When form is not at odds with communication. *Language Awareness*, *30*(1), 1–16. https://doi.org/10.1080/09658416.2020.1751178

Calzada, A., & García Mayo, M. P. (2020b). Child EFL learners' attitudes towards a collaborative writing task: An exploratory study. *Language Teaching for Young Learners*, *2*(1), 52–72. https://doi.org/10.1075/ltyl.19008.cal

Cambridge Assessment English. (2018). *Young Learners Sample Papers 2018 – Flyers A2*. Cambridge Assessment English. Retrieved from https://www.cambridgeenglish.org/Images/young-learners-sample-papers-2018-vol1.pdf

Carrell, P. (1985). Facilitating ESL reading by teaching text structure. *TESOL Quarterly*, *19*, 727–752. https://doi.org/10.2307/3586673

Collins, L., & White, J. (2019). Observing language-related episodes in intact classrooms: Context matters! In R. M. DeKeyser & G. Prieto Botana (Eds.). *Doing SLA research with implications for the classroom* (pp. 9–30). Amsterdam: John Benjamins. https://doi.org/10.1075/lllt.52.02col

Dalton-Puffer, C. (2011). Content-and-Language Integrated Learning: From practice to principles? *Annual Review of Applied Linguistics*, *31*, 182–204. https://doi.org/10.1017/S0267190511000092

Dao, P., & McDonough, K. (2017). The effect of task role on Vietnamese EFL learners' collaboration in mixed proficiency dyads. *System*, *65*, 15–24. https://doi.org/10.1016/j.system.2016.12.012

Department of Education. (2020). *ESE2 2019/2020 English literacy marking guidelines*. Government of Navarre. Retrieved from https://www.educacion.navarra.es/documents/27590/1678902/ESO2_Competencia_Ingles_19_20_CC.pdf/dcd33fbc-5b1e-4b28-478e-777128d30273

Donato, R. (1988). Beyond group: A psycholinguistic rationale for collective activity in second-language learning (Unpublished doctoral dissertation). University of Delaware, Newark.

Ellis, R., Skehan, P., Li, S., Shintani, N., & Lambert, C. (2020). *Task-based language teaching: Theory and practice*. Cambridge: Cambridge University Press.

Etxeberria, F., & Etxeberria, J. (2015). Bilingual education in the Basque Country (1960–2013). In F. Tochon (Ed.), *Language education policy unlimited: Global perspectives and local practices* (pp. 249–277). Blue Mounds, WI: Deep University Press.

Fernández Dobao, A. (2012). Collaborative writing tasks in the L2 classroom: Comparing group, pair and individual work. *Journal of Second Language Writing*, *21*(1), 40–58. https://doi.org/10.1016/j.jslw.2011.12.002

Fernández Dobao, A. (2016). Peer interaction and learning: A focus on the silent learner. In M. Sato & S. Ballinger (Eds.), *Peer interaction and second language learning: Pedagogical potential and research agenda* (pp. 33–61). Amsterdam: John Benjamins. https://doi.org/10.1075/lllt.45.02fer

Freelon, D. (2013). ReCal OIR: Ordinal, interval, and ratio intercoder reliability as a web service. *International Journal of Internet Science*, *8*(1), 10–16.

García Mayo, M. P., & Imaz Agirre, A. (2019). Task modality and pair formation method: Their impact on patterns of interaction and LREs among EFL primary school children. *System*, *80*, 165–175. https://doi.org/10.1016/j.system.2018.11.011

Guiraud, P. (1960). *Problèmes et méthodes de la statistique linguistique*. Paris: Presses universitaires de France.

Housen, A., Kuiken, F., & Vedder, I. (Eds.). (2012). *Dimensions of L2 performance and proficiency. Complexity, accuracy and fluency in SLA*. Amsterdam: John Benjamins. https://doi.org/10.1075/lllt.32

IBM Corp. (2016). *IBM SPSS Statistics for Windows. Version 24.0*. IBM Corp.

JASP Team (2019). *JASP (Version 0.11.1)* [Computer software].

Kim, Y., & McDonough, K. (2008). The effect of interlocutor proficiency on the collaborative dialogue between Korean as a second language learners. *Language Teaching Research*, *12*(2), 211–234. https://doi.org/10.1177/1362168807086288

Lantolf, J. (2012). Sociocultural theory: A dialectical approach to L2 research. In S. Gass & A. Mackey (Eds.), *Handbook of second language acquisition* (pp. 57–72). New York, NY: Routledge.

Larsen-Freeman, D. (1978). An ESL index of development. *TESOL Quarterly*, *12*(4), 439–448. https://doi.org/10.2307/3586142

Larson-Hall, J. (2016). *A guide to doing statistics in second language research using SPSS and R* (2nd ed.). New York, NY: Routledge.

Larson-Hall, J., & Herrington, R. (2010). Improving data analysis in second language acquisition by utilizing modern developments in Applied Statistics. *Applied Linguistics*, *31*(3), 368–390. https://doi.org/10.1093/applin/amp038

Lázaro-Ibarrola, A., & Hidalgo, M. Á. (2017). Procedural repetition in task-based interaction among young EFL learners: Does it make a difference? *ITL – International Journal of Applied Linguistics*, *168*(2), 183–202. https://doi.org/10.1075/itl.16024.laz

Lee, I. (2016). EFL writing in schools. In R.M. Manchón & P.K. Matsuda (Eds.). *Handbook of second and foreign language writing* (pp. 113–139). Berlin: De Gruyter. https://doi.org/10.1515/9781614511335-008

Leeser, M. (2004). Learner proficiency and focus on form during collaborative dialogue. *Language Teaching Research*, *8*(1), 55–81. https://doi.org/10.1191/1362168804lr1340a

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk* (3rd ed.) [Computer software]. Mahwah, NJ: Lawrence Erlbaum Associates.

Malmqvist, A. (2005). How does group discussion in reconstruction tasks affect written language output? *Language Awareness*, *14*(2–3), 128–141. https://doi.org/10.1080/09658410508668829

Manchón, R.M. (2011). Writing to learn the language: Issues in theory and research. In R.M. Manchón (Ed.). *Learning-to-write and writing-to-learn in an additional language* (pp. 61–82). Amsterdam: John Benjamins. https://doi.org/10.1075/lllt.31.07man

Manchón, R.M., & Matsuda, P.K. (Eds.) (2016). *Handbook of second and foreign language writing*. Berlin: De Gruyter. https://doi.org/10.1515/9781614511335

Marsden, E., Mackey A., & Plonsky, L. (2016). The IRIS Repository: Advancing research practice and methodology. In A. Mackey & E. Marsden (Eds.), *Advancing methodology and practice: The IRIS repository of instruments for research into second languages* (pp. 1-21). New York: Routledge.

Matsuda, P., & DePew, K. (2002). Early second language writing: An introduction. *Journal of Second Language Writing*, *11*, 261–268. https://doi.org/10.1016/S1060-3743(02)00087-5

McDonough, M., & García Fuentes, C. (2015). The effect of writing task and task conditions on Colombian EFL learners' language use. *TESL Canada Journal/Review TESL du Canada*, *32*(2), 67–79. https://doi.org/10.18806/tesl.v32i2.1208

Meara, P.M., & Miralpeix, I. (2017). *Tools for researching vocabulary*. Bristol: Multilingual Matters.

Michel, M. (2017). Complexity, accuracy, and fluency in L2 production. In S. Loewen & M. Sato (Eds.), *The Routledge handbook of instructed second language acquisition* (pp. 50–68). New York, NY: Routledge. https://doi.org/10.4324/9781315676968-4

Ortega, L. (2009). Studying writing across EFL contexts: Looking back and moving forward. In R.M. Manchón (Ed.). *Writing in foreign language contexts: Learning. teaching and research* (pp. 232–255). Clevedon: Multilingual Matters. https://doi.org/10.21832/9781847691859-013

Pallotti, G. (2009). CAF: Defining, refining and differentiating constructs. *Applied Linguistics*, *30*(4), 590–601. https://doi.org/10.1093/applin/amp045

Pallotti, G. (2015). A simple view of linguistic complexity. *Applied Linguistics*, *30*(4), 555–578. https://doi.org/10.1177/0267658314536435

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. https://doi.org/10.1016/j.jslw.2014.09.003

Plonsky, L., & Kim, Y. (2016). Task-based learner production: A substantive and methodological review. *Annual Review of Applied Linguistics*, *36*, 73–97. https://doi.org/10.1017/S0267190516000015

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research: Effect sizes in L2 research. *Language Learning*, *64*(4), 878–912. https://doi.org/10.1111/lang.12079

Reichelt, M., Lefkowitz, N., Rinnert, C., & Schultz, J. M. (2012). Key issues in foreign language writing. *Foreign Language Annals*, *45*(1), 22–41. https://doi.org/10.1111/j.1944-9720.2012.01166.x

Sample, E., & Michel, M. (2015). An exploratory study into trade-off effects of complexity, accuracy, and fluency on young learners' oral task repetition. *TESL Canada Journal*, *31*, 23. https://doi.org/10.18806/tesl.v31i0.1185

Shin, S-Y., Lidster, R., Sabraw, S., & Yeager, R. (2016). The effects of L2 proficiency differences in pairs on idea units in a collaborative text reconstruction task. *Language Teaching Research*, *20* (3), 366–386. https://doi.org/10.1177/1362168814567888

Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics*, *30*(4), 510–532. https://doi.org/10.1093/applin/amp047

Skehan, P. (2016). Tasks versus conditions: Two perspectives on task research and their implications for pedagogy. *Annual Review of Applied Linguistics*, *36*, 34–49. https://doi.org/10.1017/S0267190515000100

Skehan, P. (2018). *Second language task-based performance: Theory, research, assessment*. New York, NY: Routledge. https://doi.org/10.4324/9781315629766

Storch, N. (1999). Are two heads better than one? Pair work and grammatical accuracy. *System*, *27*(3), 363–374. https://doi.org/10.1016/S0346-251X(99)00031-7

Storch, N. (2002). Patterns of interaction in ESL pair work. *Language Learning*, *52*(1), 119–158. https://doi.org/10.1111/1467-9922.00179

Storch, N. (2005). Collaborative writing: Product, process and students' reflections. *Journal of Second Language Writing*, *14*, 153–173. https://doi.org/10.1016/j.jslw.2005.05.002

Storch, N. (2016). Collaborative writing. In R. M. Manchón & P. K. Matsuda (Eds.). *Handbook of second and foreign language writing*. (pp. 387–406). Berlin: De Gruyter. https://doi.org/10.1515/9781614511335-021

Storch, N. (2019). Collaborative writing. *Language Teaching*, *52*(1), 40–59. https://doi.org/10.1017/S0261444818000320

Storch, N., & Aldosari, A. (2013). Pairing learners in pair work activity. *Language Teaching Research*, *17*(1), 31–48. https://doi.org/10.1177/1362168812457530

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). Clevedon: Multilingual Matters.

Swain, M. (2006). Languaging, agency and collaboration in advanced language proficiency. In H. Byrnes (Ed.). *Advanced language learning: The contribution of Halliday and Vygotsky* (pp. 95–108). London: Continuum.

Swain, M., & Lapkin, S. (1998). Interaction and second language learning: Two adolescent French immersion students working together. *The Modern Language Journal*, *82*, 320–337. https://doi.org/10.1111/j.1540-4781.1998.tb01209.x

Swain, M., & Lapkin, S. (2001). Focus on form through collaborative dialogue: Exploring task effects. In M. Bygate, P. Skehan & M. Swain (Eds.), *Researching pedagogic tasks. Second language learning, teaching and testing* (pp. 99–117). London: Longman.

Tejada Sánchez, I., & Pérez Vidal, C. (2018). Writing performance and time of exposure in EFL immersion learners. Analysing complexity, accuracy and fluency. In C. Pérez Vidal, S. López-Serrano, J. Ament, & D. J. Thomas-Wilhelm (Eds.). *Learning context effects: Study abroad. formal instruction and international immersion classrooms* (pp. 101–129). Berlin: Language Science Press. https://doi.org/10.5281/zenodo.1446470

Tryon, W. W. (2001). Evaluating statistical difference, equivalence, and indeterminacy using inferential confidence intervals: An integrated alternative method of conducting null hypothesis statistical tests. *Psychological Methods*, *6*(4), 371–386. https://doi.org/10.1037/1082-989X.6.4.371

Villarreal, I., & Gil-Sarratea, N. (2019). The effect of collaborative writing in an EFL secondary setting, *Language Teaching Research*. https://doi.org/10.1177/1362168819829017

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. 89–115). Berlin: De Gruyter. https://doi.org/10.1515/9781501511318-005

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University 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. https://doi.org/10.1177/136216880607074599

Wajnryb, R. (1990). *Grammar dictation*. Oxford: Oxford University Press.

Yasuda, S. (2019). Children's meaning-making choices in EFL writing: The use of cohesive devices and interpersonal resources. *System*, *85*. https://doi.org/10.1016/j.system.2019.102108

## Appendix. The analytic rubric

| | | 3 | 2 | 1 |
|---|---|---|---|---|
| Task | Adequacy | All the parts of the story are included (beginning, body, ending); the length of the text is appropriate | Most parts of a story are included; the text is too short (ideas are not fully developed) | Notable omissions of the content and/or considerable irrelevance of some of them |
| | Coherence | A clear text, easy to understand | Easy to understand, although there are some incoherent points that confuse the reader | Difficult to understand |
| Language | Cohesion | Ideas are well organised (use of paragraphs). Cohesive devices linking sentences and paragraphs. No serious mistakes | Ideas are organised. Some cohesive devices linking sentences and paragraphs. There may be some mistakes | There is a lack of organisation or linking devices |
| | Grammatical accuracy | Very few, irrelevant or no grammar errors at all. Good command of grammar | Some acceptable grammar errors. Fair command of English grammar | Serious and numerous grammar mistakes |
| | Mechanics | Most words are written correctly, only some occasional mistakes | Some spelling mistakes (between 3 and 6), some of them in basic vocabulary | Many spelling mistakes. Invents words |
| | Lexical range | Rich and varied vocabulary | Basic vocabulary, enough to convey the message | Limited range of vocabulary. Some words are in the L1 |

## Address for correspondence

Asier Calzada
Centro de Investigación Micaela Portilla Ikergunea 3.6
Universidad del País Vasco
Euskal Herriko Unibertsitatea (UPV/EHU)
C/Justo Vélez de Elorriaga 1
01006 Vitoria-Gasteiz
Spain
asier.calzada@ehu.eus

## Biographical notes

**Asier Calzada** is a PhD student supervised by Professor Dr. María Pilar García Mayo at the University of the Basque Country (UPV/EHU) and a member of the *Language and Speech* research group. He has worked as an EFL teacher in private language schools in Spain and also as a lecturer in English and Russian at the University of the Basque Country. His main research interest is the use of collaborative writing tasks in primary school EFL settings. Moreover, his research also focuses on the impact of task-related variables and individual differences on children's collaborative writing performance.

**Dr. María del Pilar García Mayo** is Full Professor of English Language and Linguistics at the University of the Basque Country (UPV/EHU). She has published widely on the L2/L3 acquisition of English morphosyntax and the study of conversational interaction in EFL. She has been an invited speaker to universities in Europe, Asia and North America and is an Honorary Consultant for the Shanghai Center for Research in English Language Education. Prof. García Mayo is the director of the research group *Language and Speech* and the MA program *Language Acquisition in Multilingual Settings*. She is also the editor of *Language Teaching Research*.