Language Teaching (2025), **58**, 274–282
doi:10.1017/S0261444824000302

CAMBRIDGE UNIVERSITY PRESS

REPLICATION RESEARCH

Replication research in L2 collaborative writing: Replications of Fernández Dobao (2012) and Bikowski and Vithanage (2016)

Matt Kessler

University of South Florida, Tampa, FL, USA
Email: mattjkess@gmail.com

(Received 13 January 2024; revised 22 June 2024; accepted 23 June 2024)

Abstract

Over the past two decades in the applied linguistics subfield of second language (L2) writing, there has been considerable interest in the topic of collaborative writing (CW). Studies in this domain have investigated different phenomena such as the nature of learner-to-learner interactions, the learning outcomes of CW, and students' perceptions of these activities when implemented in the classroom. Despite the large number of studies that have been published to date, replication research has been scarce. As such, the current article opens by making a case for replication work in the area of L2 CW, arguing why such research is both important and necessary. Following this, the article turns to a discussion of two key CW studies that have been highly influential in the L2 writing sphere. These studies are described in detail, and suggestions are provided as to how and why these studies might be replicated in the future.

1. Introduction

This article presents an argument for replicating COLLABORATIVE WRITING (CW) studies within the applied linguistics subfield of second language (L2) writing. Specifically, CW refers to an activity in which two or more people are involved in the production of a written text or genre (e.g., an argumentative essay, blog post, digital poster), and in which the participants engage in all aspects of the composition process and share co-ownership of the product that is produced (Storch, 2019). During the past two decades, CW has experienced a wave of interest. To date, numerous books (e.g., Li & Zhang, 2023; Storch, 2013) and methodological syntheses (e.g., Elabdali, 2021; Storch, 2011; Zhang & Plonsky, 2020; Zhang et al., 2021) have catalogued its popularity and effectiveness as a learning activity. Additionally, there is now evidence that L2 writing teachers routinely integrate CW activities into their classroom pedagogies in a myriad of contexts (see Kessler, 2023; Kessler & Casal, 2024). As such, CW is a topic that is both important and prevalent among researchers and practitioners.

Despite the high level of activity involving CW, replication research is lacking. In fact, of the several methodological reviews, meta-analyses, and future directions pieces that have been published, only one (i.e., Zhang & Plonsky, 2020) briefly mentions the word *replicat**, yet the comment is made in reference to the transparency and accuracy of reporting within applied linguistics more broadly rather than to any existing CW studies that have been purposefully replicated. Why, then, is replication lacking in this area? Most likely, as scholars such as Mu and Matsuda (2016) have noted, the purpose of replication has been misunderstood by many linguists and L2 writing researchers, as it is typically perceived as "low prestige" or "unoriginal" (p. 202). However, such misconceptions are inaccurate, with replication research being a fundamental part of many scientific fields as a means of verifying

and confirming the findings and claims of others (Mackey, 2012). That is, as scientists, we should never assume rigor and generalizability from a single study. For fields within the social sciences like applied linguistics as well, replication is particularly important. This is because, as Porte and McManus (2019) contend, "social science deals with people, and people present us with far greater problems in the interpretation of findings from studies involving them … people are not fixed" (p. 4). Thus, replication work is vital. This, too, applies to scholarship involving L2 writing and CW.

In this piece, I argue for the need to replicate two key CW studies: Fernández Dobao (2012), which was published in the *Journal of Second Language Writing,* and Bikowski and Vithanage (2016), which was published in *Language Learning & Technology*. These pieces were selected for multiple reasons. For one, both articles have had a high impact in the domain of CW and L2 writing scholarship more broadly.[1] Second, they both share common ground in that the researchers investigated the influence of CW on text outcome measures (e.g., text quality, complexity, accuracy, and fluency (CAF)); however, each article goes about doing so in different ways, with Fernández Dobao's study focusing more so on the short-term nature of group interactions, and Bikowski and Vithanage's study focusing on the long-term effects of group interactions. Additionally, each study focuses on different target languages (Spanish in Fernández Dobao, and English in Bikowski and Vithanage). As such, both articles can be viewed as complementary, while providing different yet useful information about the effectiveness of CW.

In the sections that follow, I provide a brief background to CW research, including some of the key concepts, themes, and studies in this area. I also further explicate why replication work is needed. Then, I summarize both Bikowski and Vithanage (2016) and Fernández Dobao (2012) in considerable detail. Following each summary, specific recommendations are made in terms of how researchers might replicate them in the future. The article then closes with a brief conclusion.

2. Background

For a thorough account of CW's early history and key publications, readers are encouraged to see Storch's (2019) research timeline in *Language Teaching*. However, in terms of a brief introduction, interest in CW can be traced back to works by scholars such as Donato (1994) and Swain and Lapkin (1995, 1998); three studies that highlight how when two or more learners work together, they often engage in collaborative dialogue, and thus collectively become more capable of identifying and solving linguistic problems than when attempting to solve such problems alone. Donato's work, in particular, popularized the now ubiquitous term COLLECTIVE SCAFFOLDING – a Sociocultural Theory (SCT)-inspired concept (see Vygotsky, 1978) – whereby students move fluidly in-and-out of expert and novice roles in order to support each other's learning. Importantly, when producing output during collaborative discussions, Swain and Lapkin (1995) have argued that such discussions foster NOTICING (see Schmidt, 1990), an important process whereby learners consciously become aware of and attend to gaps in their interlanguage. In works by Swain and Lapkin (1995, 1998) and others (e.g., Storch, 1998), collaborative discussions among learners have often been referred to as LANGUAGING. For research purposes, the act of languaging has been operationalized as LANGUAGE-RELATED EPISODES (LREs), with LREs referring to instances in which learners explicitly discuss or negotiate different language-related issues (e.g., grammar, vocabulary, pronunciation). Relatedly, in addition to LREs, many researchers have approached CW from a cognitivist interactionist perspective. When doing so, such researchers have often examined CW as a site for prompting learners to engage in focus-on-form and peer feedback (see Zhang & Plonsky, 2020). Taken together, interactionist approaches, SCT, collective scaffolding, and languaging have served as the primary drivers for a sizable amount of CW scholarship.

As a result of these early works, CW research began to take flight in the early 2000s. Since that time, numerous studies have been published, which can typically be categorized into several themes. As Storch (2019) has noted, these themes include (but are not limited to) studies that examine: the various factors that influence languaging and LRE production during the CW process, such as task type,

https://doi.org/10.1017/S0261444824000302 Published online by Cambridge University Press

group sizes and dynamics, mode of interaction, and L2 learners' proficiencies (e.g., Brooks & Swain, 2009; Fernández Dobao, 2012; Kessler, 2009; Kim & McDonough, 2011; Storch, 2002; Zhang & Liu, 2023); the relationships that learners form (or do not form) during CW activities (e.g., Li & Zhu, 2013, 2017a; Storch & Aldosari, 2013); the outcomes of CW involving text-based measures (e.g., text quality, and CAF), particularly when texts are produced individually versus collaboratively (e.g., Bikowski & Vithanage, 2016; Rahimi & Fathi, 2022; Shehadeh, 2011; Storch, 2005; Villarreal & Gil-Sarratea, 2020); and, finally, studies that investigate one of the aforementioned topics in addition to learners' perceptions of the CW activities and their group interactions (e.g., Chen & Yu, 2019; Li & Zhu, 2017b; Tanrıkulu, 2020); Once again, for more on these themes, readers are encouraged to see Storch (2019). However, readers are also encouraged to review syntheses by Zhang and Plonsky (2020) and Zhang et al. (2021), along with a recent edited volume by Li and Zhang (2023).

As referenced in the introductory section, although there has been an abundance of CW literature published during the past two decades, replication work is lacking. Within L2 writing, replication studies involving other topics have demonstrated the value of revisiting existing works, particularly when it comes to addressing methodological issues and confirming and/or challenging existing findings (e.g., de Kleine & Lawton, 2018; Kessler et al., 2022; Zhang & Li, 2021). Thus, in the following section, I highlight two high-impact CW studies by Fernández Dobao (2012) and Bikowski and Vithanage (2016). As will be discussed, each study has major strengths; however, both studies also have limitations, particularly involving the methods and the dependent variable measures that were selected. Thus, owing to their influence in the CW literature, each study deserves to be revisited.

3. The original studies and suggested approaches to replication

3.1 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

This study investigated the effects of CW group size on (1) the CAF of the written output produced by L2 learners, and (2) the frequency and nature of the oral LREs students produced during their interactions. The study was motivated by SCT and by Donato's (1994) concept of collective scaffolding. Specifically, Fernández Dobao was interested in understanding what an ideal CW group size might be, as literature at the time had primarily focused on analyzing the experiences of dyads. As such, Fernández Dobao explored the experiences of intermediate Spanish foreign language learners (N = 111) as they performed the same writing task in groups of four (n = 60, 15 groups), in dyads (n = 30, 15 dyads), or individually (n = 21). Participants were enrolled in six sections of the same university-level Spanish course. Each section had a different teacher but followed the same syllabus. In the study, learners first received a 15-minute grammar review lesson on the Spanish past tense. Then, learners completed a jigsaw task in which they had to arrange 15 pictures to create a story. After arranging the pictures, students were required to produce one written text narrating the pictures (i.e., one joint story per group or dyad, or individually for the participants working alone). Participants were given 30 minutes to complete the task, and those working in groups or dyads had their oral interactions audio recorded. In total, 51 texts were collected and analyzed (15 from the groups of four, 15 from the dyads, and 21 from the individuals).

The texts were manually analyzed using 11 different CAF measures, including: number of words per clause, number of words per T-unit, and number of clauses per T-unit (syntactic complexity); mean type-token ratio (lexical complexity); grammatical errors, lexical errors, mechanical errors, error-free clauses to total clauses, error-free T-units to total T-units, and errors to words (accuracy); and total number of words produced (fluency). Inter-coder reliability was not reported for the CAF measures. Additionally, a series of non-parametric Mann-Whitney U tests were run to compare the differences among the three groups of learners; however, the checking of statistical assumptions was not reported, and it is also unclear why this particular statistical test was selected based on the study design (an issue I return to in the next subsection).

https://doi.org/10.1017/S0261444824000302 Published online by Cambridge University Press

For analyzing the LREs, Fernández Dobao drew on Swain and Lapkin's (1998) work, examining instances where students discussed or questioned their own or another's language use in some way. Transcripts of the groups' and dyads' interactions were analyzed to examine the frequency, focus, and outcome of the LREs. For focus, LREs were coded based on whether they were form-focused (examining aspects of grammar), lexis-focused (vocabulary and meaning), or mechanics-focused (pronunciation, spelling, or punctuation). LREs were also coded as to whether they were correctly resolved, incorrectly resolved, or unresolved (i.e., whether students arrived at the right/wrong answer when deliberating, or whether they could not reach a decision). In coding the LREs, two coders reported high reliability using simple agreement (92% for LRE identification, and 96% for classification). Once again, Mann-Whitney *U* tests were run to compare the differences between the LREs produced by the groups and the dyads; however, the checking of statistical assumptions was not reported prior to running this non-parametric test.

For the results of the first research question involving the CAF measures, no significant differences were found on any of the syntactic or lexical complexity measures between the groups, dyads, and individuals. However, the groups' texts were found to be more accurate than those of the dyads and the individuals on five accuracy measures, with the exception of mechanical errors. Surprisingly, no statistically significant differences were found between the dyads and the individuals on accuracy measures. In terms of fluency, no differences were found between the groups' and the dyads' texts, but individual writers produced significantly longer texts than the groups and the dyads.

In terms of LREs, although they were frequent in both group and dyad interactions during the CW activity, the groups produced significantly more LREs than the dyads. Also, groups produced significantly more mechanics-focused LREs, but there were no differences in terms of form- or lexis-focused LREs. Finally, in terms of resolution, the groups correctly resolved more LREs than the dyads. Qualitative data in the form of excerpts were also provided to showcase the nature of the LREs that occurred during learners' interactions – (this is not further discussed here). Notably, for all quantitative results for the two research questions, descriptive statistics, *p*-values, and the test statistic (*U*) were reported; however, confidence intervals (CIs) and effect sizes were not reported throughout the study.

3.1.1 Approaches to replication

Fernández Dobao's (2012) study is strong in terms of its creative design and large sample size. The findings also suggest important implications for teachers, particularly when it comes to deciding how many students to put into a group when using CW activities. However, as referenced, there are some methodological limitations. Thus, the first suggestion involves conducting an approximate replication of this study, which refers to duplicating the methods of the original study but altering some of the variables (see Porte, 2012). When doing so, one dependent variable needs to be changed, along with some of the statistical analyses. Beginning with the dependent variable, this involves the use of the mean type-token ratio for assessing lexical complexity. This change is needed because as a measure, the type-token ratio has been shown to be affected by text length (see McCarthy & Jarvis, 2010). Given the aims of Fernández Dobao's study, it makes sense that text length was not controlled for in the design, yet this also means that the type-token ratio has the potential to be problematic, particularly given the limited information that was provided surrounding its use in the study (e.g., reliability). Instead, type-token ratio might be replaced with another automated measure of lexical complexity that is not impacted by length, such as Measure of Textual Lexical Diversity (or MTLD, see Kyle et al., 2021 for more). Apart from altering this measure, when conducting this approximate replication, the only other modifications that should be made involve the selection and reporting of data analyses. Specifically, for those CAF measures that require manual coding, such as the accuracy measures, inter-coder reliability must be obtained and reported. This is because coding for some accuracy measures is notoriously challenging (see Polio & Shea, 2014). Additionally, since there are three groups that are being compared in the study, a series of *t*-tests or non-parametric Mann-Whitney *U* tests are not appropriate. This is because running multiple *t*-tests in this way introduces a greater likelihood of error. Instead, one-way ANOVAs should be run after checking and reporting the assumptions (see

https://doi.org/10.1017/S0261444824000302 Published online by Cambridge University Press

Hu & Plonsky, 2021 for more on statistical assumptions), followed by post-hoc tests to locate the source of any significant differences. Finally, when reporting the results – in addition to *M*s, *SD*s, *p*-values, and F-values – both CIs and effect sizes should be reported so that information can be gleaned about the relative impact of group size on the dependent variables (see Plonsky & Oswald, 2014 for more on effect sizes).

This approximate replication is being suggested – with only minor modifications to one dependent variable and the selection and reporting of statistical analyses – to adjust for a measure that is known to be problematic, and to apply more transparent and appropriate statistical analyses that match the study's design. Thus, an approximate replication of this nature is critical since some of the data analyses in the original study may have resulted in (in)significant findings for dependent variables such as the accuracy and lexical complexity measures.

The second suggested replication involves conducting another approximate replication. In Fernández Dobao's (2012) study, students participated in groups of four, in dyads, or individually. As some readers will notice, absent from the design are groups of three. Fernández Dobao stated that the decision to only include dyads and groups of four was made "to establish a clear difference between groups and pairs" (p. 44). Yet, since the publication of her study, other works such as Li and Zhu (2017b) have suggested that when groups are comprised of three or more students, their dynamics may vary in terms of how active all members are throughout the CW process (also see works by Fernández Dobao, 2016; Kessler, 2019). Thus, one major question is: if groups of three are integrated into the study design, do we see the same types of outcomes in terms of written text and oral LRE production? Such findings would have important implications for forming groups and, specifically, as to whether having three or four learners per group is ideal. Thus, when adding CW groups of three into the study design – because this will likely require a large N-size with many participants – an additional suggestion in this regard is to adopt a counter-balanced, repeated measures design. That is, if researchers have two sections of students to use for their study, at some point, all students could participate in the writing activities individually, in pairs, in groups of three, and in groups of four over the span of a single semester. However, among the two classes, the group formations can be counter balanced (e.g., with section A starting with learners writing individually and section B beginning with learners writing in groups of four, with each section either increasing or decreasing in size over time). This way, practice effects can be minimized in the design.

As discussed, both of these approximate replications are worthwhile endeavors. For one, they address some of the methodological limitations of the original study. Second, such modifications also have the capacity to confirm or challenge the study's findings, while also extending our collective knowledge about the formation of groups in CW activities.

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

This study examined the impact of repeated in-class CW activities on L2 English learners' development of individual writing skills, in addition to students' perceptions of either individual or CW activities. The study was motivated by SCT and previous research which suggested that CW could improve students' collective writing skills; however, Bikowski and Vithanage noted that previous studies at the time had focused on the nature of the texts produced collaboratively rather than investigating individual learning gains from CW. Therefore, the researchers conducted a quasi-experimental study with advanced-level L2 English learners ($N = 59$) who were enrolled in four sections of an English for Academic Purposes course at a university in the United States. The four sections were taught by three different instructors, but all sections followed the same curriculum. The students in these classes were divided into an experimental CW group ($n = 32$) and a control group ($n = 27$), with two sections assigned to each condition. During the 15-week semester, the experimental group completed a series of four, in-class, web-based CW essays using Google Docs (e.g., argumentative and compare-contrast essays). Students in the CW condition worked in groups of three or four and were given 45 minutes to

https://doi.org/10.1017/S0261444824000302 Published online by Cambridge University Press

complete each essay collaboratively in a computer lab. Since Google Docs is a synchronous tool, students were also given the option to work on one computer together or on different computers simultaneously. For the control group, participants also completed the same essays in class, yet they did so independently.

To assess individual learning gains, Bikowski and Vithanage adopted a pretest-posttest design. The pretest was a 30-minute compare-contrast essay (on the topic of comparing one's life now to five years ago), and the posttest was a 30-minute argumentative essay (on the topic of comparing products on the market). Both tests were rated by two raters on a 100-point analytic rubric that included sections on content, organization, grammar, and style. Inter-rater reliability using Pearson's *r* was high for the pretest ($r = 0.97$) and posttest ($r = 0.95$). Paired samples *t*-tests were then used to compare each group's gains made from pretest-to-posttest, and an independent samples *t*-test was used to compare the differences in the mean gain scores of the groups. Participants also completed a post-study online survey about their experiences, which involved Likert-scale items and open-ended items – (the qualitative items are not further discussed here). For the Likert-scale items, students in the experimental and control groups used a 5-point agreement scale to rate four statements. The statements covered: (1) how well students liked their in-class writing tasks, (2) how well students thought the tasks improved their writing skills, (3) how well students worked in groups (CW group only), and (4) how well students thought their teacher liked the writing tasks (p. 87). It is unclear what each point on the Likert scale represented, as descriptors were not explicitly stated. Independent samples *t*-tests were then used to compare the responses of the experimental and control groups. Assumptions were reported as being checked prior to running the parametric tests.

In terms of the results, paired samples *t*-tests showed that both the experimental and control groups made significant gains from pretest-to-posttest. While *M*s, *SD*s, *t*-values, and *p*-values were reported, CIs and effect sizes were not provided. Notably, in terms of comparing the two groups, an independent samples *t*-test revealed that the CW group made larger, statistically significant gains when compared with the individual writing group, with a moderate effect size ($d = 0.58$). Finally, when it came to students' perceptions of the in-class writing activities, Bikowski and Vithanage reported that there were no statistically significant differences between the responses given by the two groups. The authors did note that no CW group participants provided negative evaluations of the activities when rating statements; however, in their qualitative comments, some students did voice a number of concerns (e.g., finding it stressful to merge ideas, disliking the time pressure of the task, and preferring teacher feedback over that of their peers).

3.2.1 Approaches to replication

Similar to Fernández Dobao (2012), Bikowski and Vithanage's (2016) study has a solid sample size (see Loewen & Hui, 2021 for more on sample sizes in SLA). It is also unique in its design, being one of the first attempts to examine the longitudinal effects of CW activities on individual learning gains. Despite this, their study could (and should) be replicated in multiple ways. The first suggested replication involves conducting a close replication. Specifically, one area of the design that might be modified involves the number of teachers used in the study. As Bikowski and Vithanage also acknowledged, this constituted a limitation in their study, in that the sections that comprised the control group were taught by one instructor, while the experimental group sections were taught by two different instructors (see p. 83). Fernández Dobao's study also had multiple teachers involved; however, her focus was on the impact of group interactions at a singular point in time, making it unlikely that the teachers' instructional practices came into play in determining the outcomes of learners' interactions in an isolated task with no teacher involvement. Yet, in Bikowski and Vithanage's study, the researchers are essentially examining (1) the impact of repeated in-class CW activities, and (2) the impact of instruction over time, even though this is an unnamed variable in their study. Thus, given that not all instructors are equally as skilled or motivating to their students when it comes to L2 teaching (e.g., Li, 2022), it is important to replicate Bikowski and Vithanage's study by controlling for this variable of instruction. A close replication could be conducted with the singular modification

https://doi.org/10.1017/S0261444824000302 Published online by Cambridge University Press

of having both the control group and the experimental group taught by the same instructor. Conducting such a replication would be important for eliminating this potentially confounding variable of instruction, and thus confirming or challenging the original study's findings.

The second suggested replication also involves another close replication. As was noted in the earlier description of Bikowski and Vithanage's (2016) study, to assess learners' perceptions of engaging in either individual or group writing assignments, the researchers used a 5-point Likert scale of agreement, in which participants rated four statements. Independent samples *t*-tests were then used to compare the responses of the two groups, but no significant differences were detected. Notably, while there are a range of point options available for Likert scales of agreement (i.e., 5-, 6-, 7-points, etc.), when it comes to assessing user experience and perceptions, there is evidence that 7-point scales may be preferrable. As Finstad (2010) and others have noted, using a 7-point Likert scale in such a scenario may be more appropriate than a 5-point scale. This is because 7-point scales tend to be more sensitive and accurate in assessing participants' evaluations of a topic. Therefore, because Bikowski and Vithanage used a smaller scale, this may not have allowed for enough discrimination among the items, and thus potentially resulted in insignificant findings when comparing the two groups' experiences. As such, a close replication of the authors' original study could be undertaken in which a 7-point Likert scale is used to assess learners' perceptions, with the following scale descriptors: 7 = *strongly agree*, 6 = *agree*, 5 = *somewhat agree*, 4 = *neutral*, 3 = *somewhat disagree*, 2 = *disagree*, and 1 = *strongly disagree*. Relatedly, although no CW group participants provided negative evaluations of the activities when rating statements, several students did voice concerns in their written (qualitative) comments. As such, in addition to expanding the points on the Likert-scale, future studies might consider expanding the number of statements that the students rate. For example, apart from those four statements used in the original study, more questions might be posed about the students' perceptions of the CW activity, including some of those qualitative themes identified by Bikowski and Vithanage (e.g., the ease of merging ideas, the suitability of the time allotted for the activity, and the students' preferences for teacher versus peer feedback).

Each suggested replication is important in different ways. First, by controlling for the independent variable of instruction, this has the capacity to eliminate any doubt about the nature of a potentially confounding variable. Second, by using an expanded scale in the second suggested replication, this has the capacity to detect potential differences among the groups that may not have been captured in the original study. Bikowski and Vithanage's (2016) study is frequently cited as evidence of the long-term effects of CW. Thus, replications are necessary in order to confirm the effects of repeated CW activities on individual gains and whether or not students perceive such activities as being more/less beneficial than solitary writing for their L2 learning over time.

4. Conclusion

In this article, I have attempted to make the case that replication work is not only worthwhile, but that it is also vital for researchers who are interested in CW and L2 writing more broadly. In particular, I have proposed that two key, high-impact studies be replicated in a variety of ways, primarily by attending to and addressing some of their methodological limitations. I also note that the suggestions here are not intended to criticize or call into question the works of others, but instead, to raise important questions and to engage in practices that will contribute to the generalizability of our collective knowledge. Thus, I hope that some readers will consider replicating these studies in the future, as a means of either confirming or challenging these studies' findings, particularly since they suggest important implications for L2 writing teachers in terms of organizing CW activities in their classrooms. Finally, I hope that readers will consider conducting replication studies beyond what is proposed here, both in L2 writing and in other domains.

Note

[1] As of September 2024 – when this article was finalized – according to Google Scholar, Fernández Dobao (2012) had been cited 800 times and Bikowski and Vithanage (2016) had been cited 285 times.

https://doi.org/10.1017/S0261444824000302 Published online by Cambridge University Press

References

Bikowski, D., & Vithanage, R. (2016). Effects of web-based collaborative writing on individual L2 writing development. *Language Learning & Technology*, *20*(1), 79–99. http://hdl.handle.net/10125/44447

Brooks, L., & Swain, M. (2009). Languaging in collaborative writing: Creation and response to expertise. In A. Mackey & C. Polio (Eds.), *Multiple perspectives on interaction in SLA* (pp. 58–89). Lawrence Erlbaum.

Chen, W., & Yu, S. (2019). A longitudinal case study of changes in students' attitudes, participation, and learning in collaborative writing. *System*, *82*, 83–96. doi:10.1016/j.system.2019.03.005

de Kleine, C., & Lawton, R. (2018). An analysis of grammatical patterns in generation 1.5, L1 and L2 students' writings: A replication study. *Journal of Second Language Writing*, *42*, 12–24. doi:10.1016/j.jslw.2018.10.003

Donato, R. (1994). Collective scaffolding in second language learning. In J. P. Lantolf & G. Appel (Eds.), *Vygotskian approaches to second language research* (pp. 33–56). Ablex.

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*, 100788. doi:10.1016/j.jslw.2020.100788

Fernández Dobao, A. (2012). Collaborative writing tasks in the L2 classroom: Comparing group, pair, and individual work. *Journal of Second Language Writing*, *21*(5), 40–58. doi: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). John Benjamins.

Finstad, K. (2010). Response interpolation and scale sensitivity: Evidence against 5-point scales. *Journal of Usability Studies*, *5*(3), 104–110.

Hu, Y., & Plonsky, L. (2021). Statistical assumptions in L2 research: A systematic review. *Second Language Research*, *37*(1), 171–184. doi:10.1177/0267658319877433

Kessler, G. (2009). Student-initiated attention to form in wiki-based collaborative writing. *Language Learning and Technology*, *13*(1), 79–95. http://hdl.handle.net/10125/44169

Kessler, M. (2019). Promoting text co-ownership and peer interactions in collaborative writing. *TESOL Journal*, *11*(2), e476. doi:10.1002/tesj.476

Kessler, M. (2023). Designing collaborative writing tasks for face-to-face and computer-mediated communication contexts. In M. Li & M. Zhang (Eds.), *L2 collaborative writing in diverse learning contexts* (pp. 184–201). John Benjamins.

Kessler, M., & Casal, J. E. (2024). English writing instructors' use of theories, genres, and activities: A survey of teachers' beliefs and practices. *Journal of English for Academic Purposes*, *69*, 101384. doi:10.1016/j.jeap.2024.101384

Kessler, M., Ma, W., & Solheim, I. (2022). The effects of topic familiarity on text quality, complexity, accuracy, and fluency: A conceptual replication. *TESOL Quarterly*, *56*(4), 1163–1190. doi:10.1002/tesq.3096

Kim, Y., & McDonough, K. (2011). Using pretask modelling to encourage collaborative language learning opportunities. *Language Teaching Research*, *15*(2), 183–199. doi:10.1177/1362168810388711

Kyle, K., Crossley, S. A., & Jarvis, S. (2021). Assessing the validity of lexical diversity indices using direct judgements. *Language Assessment Quarterly*, *18*(2), 154–170. doi:10.1080/15434303.2020.1844205

Li, C. (2022). Foreign language learning boredom and enjoyment: The effects of learner variables and teacher variables. *Language Teaching Research*. (Online first). doi:10.1177/13621688221090324

Li, M., & Zhang, M. (Eds.). (2023). *L2 collaborative writing in diverse learning contexts*. John Benjamins.

Li, M., & Zhu, W. (2013). Patterns of computer-mediated interaction in small writing groups using wikis. *Computer Assisted Language Learning*, *26*(1), 61–82. doi:10.1080/09588221.2011.631142

Li, M., & Zhu, W. (2017a). Good or bad collaborative wiki writing: Exploring links between group interactions and writing products. *Journal of Second Language Writing*, *35*(1), 38–53. doi:10.1016/j.jslw.2017.01.003

Li, M., & Zhu, W. (2017b). Explaining dynamic interactions in wiki-based collaborative writing. *Language Learning & Technology*, *21*(2), 96–120. http://hdl.handle.net/10125/44613

Loewen, S., & Hui, B. (2021). Small samples in instructed second language acquisition research. *Modern Language Journal*, *105*(1), 187–193. doi:10.1111/modl.12700

Mackey, A. (2012). Why (or why not), when, and how to replicate research. In G. Porte (Ed.), *Replication research in applied linguistics* (pp. 21–46). Cambridge.

McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, *42*, 381–392. doi:10.3758/BRM.42.2.381

Mu, C., & Matsuda, P. (2016). Replication in L2 writing research: *Journal of Second Language Writing* author's perceptions. *TESOL Quarterly*, *50*(1), 201–219. doi:10.1002/tesq.284

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

Polio, C., & Shea, M. C. (2014). As investigation into current measures of linguistic accuracy in second language writing research. *Journal of Second Language Writing*, *26*, 10–27. doi:10.1016/j.jslw.2014.09.003

Porte, G. (Ed.). (2012). *Replication research in applied linguistics*. Cambridge.

Porte, G., & McManus, K. (2019). *Doing replication research in applied linguistics*. Routledge.

Rahimi, M., & Fathi, J. (2022). Exploring the impact of wiki-mediated collaborative writing on EFL students' writing performance, writing self-regulation, and writing self-efficacy: A mixed methods study. *Computer Assisted Language Learning*, *35*(9), 2627–2674. doi:10.1080/09588221.2021.1888753

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

Shehadeh, A. (2011). Effects and student perceptions of collaborative writing in L2. *Journal of Second Language Writing*, *20*(4), 286–305. doi:10.1016/j.jslw.2011.05.010

Storch, N. (1998). Comparing second language learners' attention to form across tasks. *Language Awareness*, *7*(4), 176–191. doi:10.1080/09658419808667108

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

Storch, N. (2005). Collaborative writing: Product, process, and students' reflections. *Journal of Second Language Writing*, *14*(3), 153–173. doi:10.1016/j.jslw.2005.05.002

Storch, N. (2011). Collaborative writing in L2 contexts: Processes, outcomes, and future directions. *Annual Review of Applied Linguistics*, *31*, 275–288. doi:10.1017/S0267190511000079

Storch, N. (2013). *Collaborative writing in L2 classrooms*. Multilingual Matters.

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

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

Swain, M., & Lapkin, S. (1995). Problems in output and the cognitive processes they generate: A step towards second language learning. *Applied Linguistics*, *16*(3), 371–391. doi:10.1093/applin/16.3.371

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

Tanrıkulu, F. (2020). Students' perceptions about the effects of collaborative digital storytelling on writing skills. *Computer Assisted Language Learning*, *35*(5), 1090–1105. doi:10.1080/09588221.2020.1774611

Villarreal, I., & Gil-Sarratea, N. (2020). The effect of collaborative writing in an EFL secondary setting. *Language Teaching Research*, *24*(6), 874–897. doi:10.1177/1362168819829017

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.

Zhang, M., Gibbons, J., & Li, M. (2021). Computer-mediated collaborative writing in L2 classrooms: A systematic review. *Journal of Second Language Writing*, *54*, 100854. doi:10.1016/j.jslw.2021.100854

Zhang, M., & Liu, Q. (2023). Synchronous and asynchronous online collaborative writing: A study on Chinese language learners. *Foreign Language Annals*, *56*(3), 740–763. doi:10.1111/flan.12704

Zhang, M., & Plonsky, L. (2020). Collaborative writing in face-to-face settings: A substantive and methodological review. *Journal of Second Language Writing*, *49*, 100753. doi:10.1016/j.jslw.2020.100753

Zhang, X., & Li, W. (2021). Effects of n-grams on the rated L2 writing quality of expository essays: A conceptual replication and extension. *System*, *97*, 102437. doi:10.1016/j.system.2020.102437

Matt Kessler is an Assistant Professor of Applied Linguistics at the University of South Florida, USA. His research focuses on topics related to L2 writing and computer-assisted language learning. He currently serves as the co-editor of *TESOL Quarterly*'s Brief Reports section. He is also the author of *Digital multimodal composing: Connecting theory, research and practice in second language acquisition* (Multilingual Matters) and the co-author of *Making the most of graduate school: A practical guidebook for students in applied linguistics, education, and TESOL* (Applied Linguistics Press).

Cite this article: Kessler, M. (2025). Replication research in L2 collaborative writing: Replications of Fernández Dobao (2012) and Bikowski and Vithanage (2016). *Language Teaching*, 58(2), 274–282. https://doi.org/10.1017/S0261444824000302