Journal of Second Language Writing 52 (2021) 100788

Contents lists available at ScienceDirect

Journal of Second Language Writing

journal homepage: www.elsevier.com/locate/jslw

Are two heads really better than one? A meta-analysis of the L2 learning benefits of collaborative writing

Rima Elabdali [1]

Linguistics Department, Georgetown University, Box 571051, Washington, DC, 20057-1051, USA

ARTICLE INFO

Keywords:
Collaborative writing
L2 writing
Meta-analysis
Effect sizes
CAF measures
Rubric scores

ABSTRACT

The benefits of collaborative writing (CW) for second language (L2) learning rest on solid theoretical underpinnings. Unequivocal empirical evidence of those benefits has proven difficult, however. This meta-analysis examines the accumulated evidence by synthesizing the results of 33 studies that explored the learning effects of CW and by gauging the magnitude of the difference between collaborative and individual writing. The characteristics and research practices of these product-oriented CW studies were also synthesized. Across studies, the effects of CW have been assessed through complexity, accuracy, and fluency (CAF) measures as well as rubric scores and grammar/vocabulary tests. The accumulated findings suggest that collaboratively written texts were more accurate than individually written texts, with a mean effect size of a medium magnitude (g=0.73). Further, a large magnitude difference in rubric scores (g=0.94) was found in favor of individual texts written after experimental CW conditions compared to those written after control individual writing conditions. Findings regarding other dependent variables were inconclusive because of insufficient accumulation and excessive variability across individual studies. Methodological and pedagogical implications are discussed.

1. Introduction

Collaborative writing (CW) has gained attention among researchers of second language (L2) writing during the last three decades (Ede & Lunsford, 1990; Li, 2018; Storch, 2019; Zhang & Plonsky, 2020). In CW tasks, which can be completed face-to-face or online, writers co-author single texts by collaboratively generating content, negotiating structure, and revising drafts (Storch, 2013). Support for CW stems from cognitive and sociocultural theories that emphasize interaction as an opportunity to negotiate meaning, notice gaps, and receive feedback (Gass & Mackay, 2007; Long, 1996), as well as from the hypothesis that collaboration affords a social context to co-construct knowledge, pool expertise, and engage in collective scaffolding (Donato, 1994; Lantolf & Thorne, 2007).

In a review of 94 quantitative primary studies conducted between 1992 and 2017 and published ahead of print after the present study was conducted, Zhang and Plonsky (2020) synthesized methodological practices in CW as a research domain. While they limited their synthesis to face-to-face CW, they allowed a diversity of designs into their review. Two strands of research can be distinguished within the pool of studies they synthesized. One involves the writing process, with a focus on learners' discussions, revisions, and interaction patterns (Arnold, Ducate, & Kost, 2012; Li & Zhu, 2017; Storch & Aldosari, 2013). The other concerns product-oriented studies which investigate the benefits of CW by examining texts (i.e., products). In the present study, I meta-analyze the substantive findings of the second strand of research, the product-oriented studies of CW, and include studies in both face-to-face and online

E-mail address: rme34@georgetown.edu.
[1] Present Address: Linguistics Department, Georgetown University, Box 571051, Washington, DC, 20057-1051. USA.

https://doi.org/10.1016/j.jslw.2020.100788
Received 16 July 2020; Received in revised form 7 November 2020; Accepted 11 November 2020
Available online 25 January 2021
1060-3743/© 2020 Elsevier Inc. All rights reserved.

modes. Within product-oriented studies research designs can vary too: Some compare collaborative and individual texts composed within a single session (e.g., Fernández Dobao, 2012; Villarreal & Gil-Sarratea, 2019), while others compare CW to individual writing using a variety of posttest measures (e.g., Hsu & Lo, 2018; Shehadeh, 2011).

The product-oriented studies of CW to date have yielded mixed results (Storch, 2018), making it difficult to draw firm conclusions on the learning potential of CW tasks. Thus, the goal of the present meta-analysis is to offer a fine-grained analysis of the purported benefits of CW for L2 learning (e.g., accuracy, fluency, complexity) and L2 writing (i.e., text quality) as well as inform L2 writing instruction with the accumulated findings of collaborative writing studies conducted in instructional contexts thus far. The chosen method of meta-analysis is ideal to reveal the current state of substantive knowledge about CW as a treatment, while also revealing areas for improved empirical practices (Oswald & Plonsky, 2010). I address these issues in the present meta-analysis of 33 product-oriented CW studies published between 2002 and March 2019 by synthesizing main study characteristics and research practices, and by analyzing effect sizes that represent the magnitude of difference between collaborative and individual writing.

2. Product-oriented collaborative writing research

Product-oriented CW research has evolved into a domain addressing two main research questions: 1) how do texts written collaboratively compare to texts written individually? And 2) how do participants who experience CW as a treatment compare to participants who only experience individual writing (i.e., control) on individual posttests of writing, grammar, or vocabulary? The first question has been pursued in what I will call one-shot design studies, whereas the second question has been addressed in experimental designs. In what follows, I discuss the two research designs in turn.

2.1. One-shot design studies

To examine the effects of collaboration on textual output, researchers have employed a one-shot design that includes a group engaged in (face-to-face or computer-mediated) CW and a group engaged in individual writing. The composed texts are then compared using a variety of dependent variables (e.g., complexity, accuracy, fluency, and rubric scores). The underlying logic is that learners' pooling of linguistic resources during the collaboration process should result in higher scores for the collaborative texts.

The findings from one-shot product-oriented studies are mixed. An early study by Storch (2005) compared the quality of face-to-face collaboratively written and individually written texts and reported no statistically significant differences in measures of complexity, accuracy, and fluency (CAF). However, other studies showed a statistically significant positive advantage for the CW condition on accuracy measures (e.g., Fernández Dobao, 2012; McDonough, De Vleeschauwer, & Crawford, 2018; Wigglesworth & Storch, 2009) but not on fluency, complexity, or rubric scores, with some studies even showing positive effects in favor of individual writing (e.g., for fluency in Fernández Dobao, 2012; and for complexity in McDonough et al., 2018). Similar inconsistencies have been observed in computer-mediated studies. Elola and Oskoz (2010) reported no statistically significant differences in CAF measures, whereas other studies noted statistically significant effects in favor of collaboration on fluency (Liou & Lee, 2011; Pae, 2011; Strobl, 2014), lexical complexity (Pae, 2011), and on content selection and organization (Strobl, 2014). Still, no effects were observed for other measures employed in the very same studies such as accuracy, syntactic complexity, cohesion, or overall rubric scores.

Overall, it is difficult to conclude definitively that the theoretically motivated superiority of collaborative texts compared to individual texts is empirically supported when using one-shot designs. It should be noted that in most of these studies, the collaborative and individual texts compared are written by different writers (i.e., a between-groups design, (e.g., Villarreal & Gil-Sarratea, 2019). Nevertheless, in some studies, the same writers served as their own control by writing one text collaboratively and another individually (i.e., a within-groups design, (e.g., Stell, 2018). It is unclear, however, whether the conflicting results might be related to the between-groups vs. within-groups designs employed across one-shot studies.

2.2. Experimental design studies

Even if it were possible to demonstrate that CW results in superior texts than individual writing in one-shot, direct comparisons, researchers would still want to see evidence that the benefits of writing collaboratively can endure beyond one single writing experience. To examine the effects of collaboration on subsequent individual performance, researchers have employed an experimental design in which two groups (i.e., experimental and control) are first pretested to ensure an equivalent baseline. For the treatment, the experimental group engages in (face-to-face or computer-mediated) CW while the control group engages in individual writing. Finally, learners in both groups complete individual posttests, and those posttests are compared to inspect the effects of the treatment. Experimental design studies have employed two types of posttests: (1) writing posttests and, less frequently, (2) grammar/vocabulary posttests. When writing posttests are employed, researchers often employ analytic rubrics that include content, organization, grammar, vocabulary, and mechanics (e.g., Khatib & Meihami, 2015). In the few studies that utilized grammar/vocabulary posttests, writing tasks are seeded with target grammar or vocabulary to examine whether the treatment can lead to learners' acquisition of those structures (e.g., Alammar, 2017; Kim, 2008).

The findings gleaned from studies featuring an experimental design suggest that the benefits of CW on subsequent individual writing can endure in terms of content and vocabulary (e.g., Shehadeh, 2011). However, for organization, grammar, and mechanics, the hypothesized subsequent benefits are not consistently seen. For example, results from studies featuring grammar/vocabulary posttests are mixed, with some finding an advantage for the collaborative group in vocabulary (e.g., Kim, 2008; Liu & Lan, 2016) and grammar (e.g., Alammar, 2017; Reinders, 2009) while others (e.g., Kuiken & Vedder, 2002) reporting no such gains.

3. The present study

The narrative review of product-oriented CW studies just presented clearly shows that evidence for the effectiveness of L2 CW is mixed, and that this applies to both one-shot and experimental research designs. Several reasons can be posited to explain the lack of consistency. As Manchón (2011) maintained and Zhang and Plonsky (2020) seconded, the nature and magnitude of learning in CW tasks are likely mediated by learner and task-related factors, including L2 proficiency (e.g., Storch & Aldosari, 2013), task type (e.g., Storch & Wigglesworth, 2007), task mode (e.g., Wang, 2015) and number of collaborators (e.g., Bueno Alastuey & Martínez de Lizarrondo Larumbe, 2017). However, before the field can systematically address these complex interactions, it is crucial to evaluate the accumulated evidence more systematically. In the absence of consistent answers about the overall effectiveness of CW, there is no cumulative context for situating efforts at analyzing interactions among variables, nor for charting new research directions in this domain. The synthetic reviews of L2 CW that exist (e.g., Li, 2018; Storch, 2019; Zhang & Plonsky, 2020) have not attempted to meta-analyze quantitative outcomes. The current study, therefore, examined the state of cumulative knowledge within product-oriented CW research by synthesizing primary research studies for key study features and research practices, and by extracting effect sizes and interpreting meta-analytic findings. Specifically, the study addressed the following questions:

1 How have the language learning outcomes of CW been defined/measured in product-oriented CW studies?
2 How do texts written collaboratively compare to texts written individually? That is, what are the accumulated findings from one-shot CW designs?
3 How do participants who experience CW as a treatment compare to participants who only experience individual writing (i.e., control) on individual posttests? That is, what are the accumulated findings from experimental CW designs?

3.1. Search and retrieval procedures

The research domain was defined as sources discussing CW tasks in language learning contexts (i.e., L2 CW). The studies were located through the systematic search procedures depicted in Fig. 1. I first created a repository of keywords by examining studies reviewed in Storch (2019). As suggested by Plonsky and Brown (2015), the search included the six databases shown in Fig. 1. The keyword search was restricted to the documents' titles, abstracts, and keywords. The time frame was not specified, but the publication language was restricted to English because the available resources did not allow for accurately representing studies in other languages. The search resulted in 2,795 potentially relevant studies. To eliminate false hits and duplicates, I examined titles, abstracts, and occasionally full reports, reducing the number to 254 studies. From these studies, I created a list of 12 journals that frequently publish

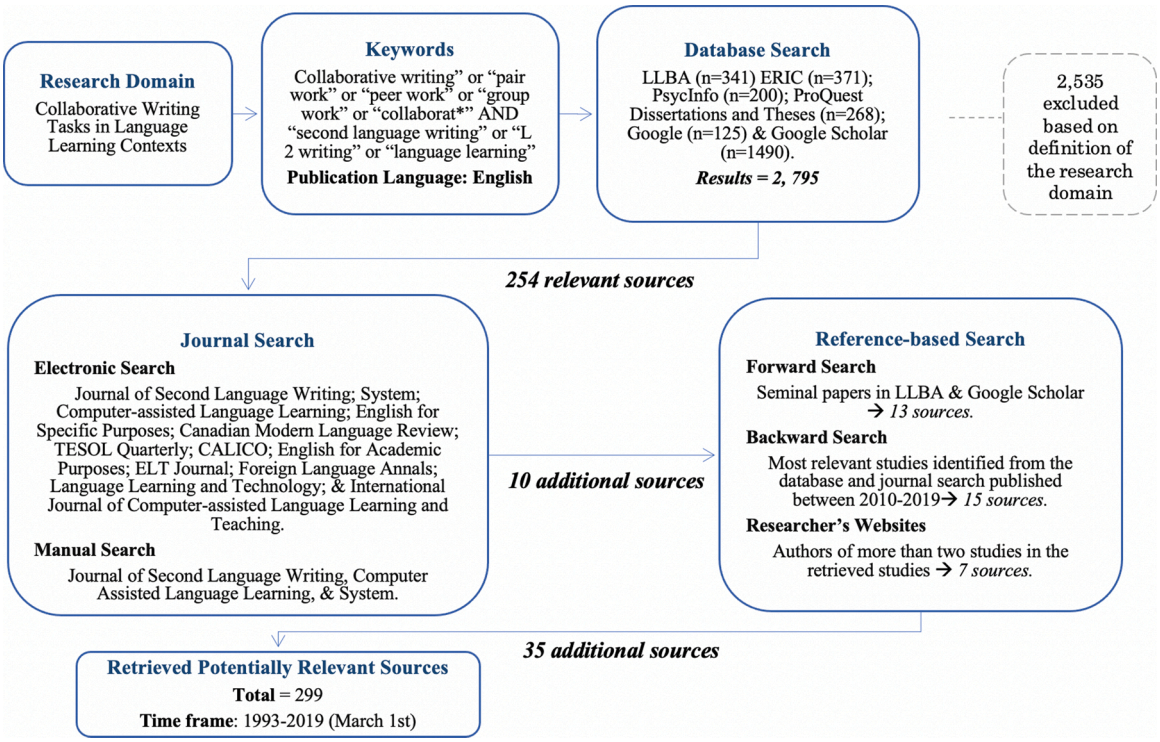

Fig. 1. Search Procedures and Results.

research on CW and electronically searched them. I also performed a manual search in three journals that published most retrieved studies. The search process also included backward and forward searches. Finally, I searched the websites of researchers who authored more than two retrieved studies. The search process generated a list of 299 CW studies that were considered potentially relevant, ranging from 1992 (i.e., earliest retrieved study) to studies published prior to March 1st, 2019. These 299 studies were inspected in more detail and the inclusion and exclusion criteria were created.

3.2. Inclusion and exclusion criteria

The studies included in the current meta-analysis met the following criteria:

1. The study analyzed empirical primary data.
2. The study was reported as a journal article, book chapter, or doctoral dissertation. MA theses, conference proceedings, and unpublished manuscripts were not considered because they are difficult to retrieve exhaustively, which may introduce unknown biases.
3. The collaborative writing task involved composing a text (e.g., essay), or compositions based on short dictations (i.e., dictogloss tasks), pictorial prompts (e.g., jigsaw), or reading passages. Tasks that require learners to collaborate by only inserting missing words or correcting errors are fundamentally different from composing tasks (Storch, 2013).
4. The product of the CW task was a single co-authored text, rather than each participant producing their own separate text upon collaborating, because joint ownership of texts is a key element of the CW experience (Storch, 2018).
5. The writing took place in either the traditional (face-to-face) mode or online.

In addition, one or more of the following reasons led to the exclusion of a study from the meta-analysis:

1. The study was process-oriented, that is, it examined learners' interactions and/or their perceptions of CW.
2. The study was product-oriented but (a) failed to include a control group completing an individual writing task or (b) compared outcomes from different CW groups without an individual writing comparison.
3. One of the collaborators was an L1 speaker or a teacher.
4. Studies that did not report data needed to calculate effect sizes such as the number of participants (*N*), mean (*M*), Standard deviation (*SD*), t-tests, or ANOVAs, if efforts to obtain the data from the authors were unsuccessful.

A full explanation of the exclusion criteria, the number of studies excluded for each criterion, and examples of each can be found in the supplemental files. After applying the criteria, 33 studies were included and 266 were excluded from the final meta-analysis.

3.3. Coding scheme

A coding scheme, available in the supplemental files, was developed by consulting previous reviews (e.g. Storch, 2013) and later expanded by closely examining around 10 of the included studies (30%) in order to identify candidates for new coding categories. This process allowed room for critical conceptual consideration of the final codlings. For example, for task type, instead of the traditional grammar-focused vs. meaning-focused categorization, and based on the tasks featured in the 30 % initial studies consulted, I categorized tasks into content-reproducing tasks, which provide learners with content to be conveyed when writing (i.e., dictogloss, data commentary, jigsaw, and summary writing), and content-generating tasks which require learners to develop their own content based only on written prompts (i.e., paragraphs and essays). This categorization is more descriptive and does not make any assumptions about learners' orientation to grammar or meaning. The final coding scheme was used to extract four main categories about each study: a) context and participants, b) study design and variables, c) task design, d) statistical data to calculate effect sizes. An additional linguist trained in meta-analyses coded 6 studies (20%), yielding a sufficient agreement among raters (inter-rater agreement ranged from 93.11%-95.23% depending on individual items; Cohen's kappa = .90). Based on this initial coding, disagreements were discussed and resolved, and definitions in the coding book were refined. I then coded the remainder of 27 studies. (see supplemental files for coding results)

3.4. Effect sizes within studies

Following previous suggestions (e.g., Oswald & Plonsky, 2010), Cohen's *d* was used to calculate effect sizes for contrasts between CW (experimental group/condition) and individual writing (control group/condition) outcomes within each primary study. When possible, Cohen's *d* was calculated based on means and standard deviations (see supplemental files for a more detailed account of effect size calculations). Since Cohen's *d* may overestimate effect sizes in small samples (Borenstein, Hedges, Higgins, & Rothstein, 2011), I also calculated the unbiased estimate, Hedges' *g*. To estimate statistical trustworthiness, I calculated the standard error and 95% confidence intervals (CIs) for each effect size using individuals within the studies as sampling units. CIs are interpreted in terms of width; narrower CIs are considered robust and those that fall above zero indicate that the mean effect size is trustworthy (Norris & Ortega, 2000). As measures of trustworthiness, CIs are akin to traditional statistical significance testing and are susceptible to sample size (i.e., the more observations the narrower the CIs).

To determine appropriate contrasts for effect size calculations, the following procedure was employed:

1 When multiple measures were used to investigate the same construct underlying the same dependent variable in a study, I averaged their effect sizes to create one synthetic effect size. For CAF measures of "accuracy," this meant averaging percentage of error free T-units and percentage of error free clauses in the same study. For rubrics, this meant averaging effect sizes for subscale ratings (e.g., content, organization, and grammar) if composite rubric scores were not reported.
2 For studies that examined multiple constructs as dependent variables (e.g., CAF), effect sizes were not averaged into a single one. The effect sizes from, say, accuracy versus complexity were likely associated in theoretically meaningful ways, for example, in a trade-off relation (Skehan, 2003). Averaging interdependent effect sizes may underestimate the heterogeneity and "delimit or confound the actual range of findings reported in primary research" (Norris & Ortega, 2000, p. 448).
3 For experimental design studies, which reported pretest and posttest mean scores for both the control and experimental groups, I calculated effect sizes on both the pretest and posttest. I then calculated adjusted effect sizes by subtracting pretest effect sizes from posttest effect sizes. This procedure was employed because relying on posttest between-group effect sizes could inflate the magnitude (Durlak, 2009).
4 In three primary studies that included one control group and more than one experimental group (e.g., pairs and triads), I averaged the data from the two experimental groups to preserve the assumption of independence (Borenstein et al., 2011). I calculated the weighted mean and the combined standard deviation for the two groups and used those to calculate the effect size and variance.

3.5. Mean effect sizes across studies

Two main issues arose when combine individual study effect sizes into averages. These are discussed in detail in the supplemental files and summarized here. First, the mean effect sizes for one-shot and experimental designs were not combined in order to preserve conceptual clarity about what each design entails. However, as mentioned, one-shot studies can compare collaborative and individual texts by either different writers (i.e., a between-groups design, $N = 12$ in the present study) or the same writers (i.e., a within-groups design, $N = 6$). Calculation of mean effect sizes comparing two different groups versus pre-to-post effect sizes of the same group have typically been kept separate in other meta-analyses (e.g., Lee, Jang, & Plonsky, 2014; Plonsky & Zhuang, 2019), for good reasons: 1) treatment effects in within-group designs are confounded with practice effects, and 2) a smaller standard deviation in within-group designs can inflate average effect sizes. However, combining these two kinds of effect sizes in instructed SLA is not unheard of (for a recent example, see McAndrews, 2019). In the present case, the low number of accumulated primary studies precluded the estimation of a trustworthy average for each design separately. Therefore, I decided to combine effect sizes from between-group and within-group one-shot design studies, for two reasons. First, all six within-group studies included in the current meta-analysis implemented a counterbalanced design to offset practice effects. Moreover, when I conducted a sensitivity analysis by disaggregating the combined average into two averages representing the two designs, the CIs for the two averages overlapped significantly, indicating that the two designs were not substantially different.

Using a random effects model (Borenstein et al., 2011), I calculated weighted mean effect sizes for four dependent variables examined in one-shot design studies (i.e., accuracy, complexity, fluency, and rubric scores) and two dependent variables in experimental design studies (i.e., rubric scores and test scores). The standard error and CIs were calculated for each mean effect size. It should be noted that the low number of studies in the meta-analysis ($N = 33$) precluded an estimation of publication bias, since publication bias tests have low statistical power when the number of effect sizes for aggregation is small (van Aert, Wicherts, & van Assen, 2019). Likewise, the low number of studies renders comparisons among different moderator variables too unstable for trustworthy interpretations. In lieu of moderator analyses, a close inspection of the individual study results will be pursued so as to offer tentative suggestions for moderators that might be worthwhile meta-analyzing in the future.

4. Results

The results of the coding and statistical analyses are presented in this section. I start by describing the database characteristics, which is then followed by findings addressing the three research questions.

4.1. Characteristics of the database

The final database consisted of 33 studies published between 2002 and 2019, including 28 journal articles and five dissertations (marked with an asterisk in the references). Table 1 summarizes the characteristics of the studies. The majority (72 %) focused on

Table 1
Database Characteristics.

Publication Type	Journal article ($N = 28$); Dissertation ($N = 5$)
Research Context	EFL ($N = 24$); ESL ($N = 5$); FL ($N = 3$); SL ($N = 1$)
Instructional Status	University ($N = 26$); Language institute ($N = 3$); High school ($N = 3$); Middle school ($N = 1$)
Proficiency Level	Intermediate ($N = 19$); Beginners ($N = 3$); Advanced ($N = 3$); Mixed ($N = 3$); Not reported ($N = 5$)
Proficiency-level Criteria	Institutional status ($N = 15$); In-house assessment ($N = 10$); Standardized test ($N = 7$); Impressionistic judgement ($n = 1$)
Task Type	Content-generating ($N = 18$); Content-reproducing ($N = 15$)
Task Mode	Face to face ($N = 25$); Computer-mediated ($N = 7$); Mixed ($N = 1$)

English as a foreign language (EFL) contexts with participants sharing the same first language (L1). Five studies focused on English as a second language (ESL) contexts with participants from mixed L1 backgrounds. Only four studies examined other target languages, three in foreign language (FL) contexts and one in a second language (SL) context. Overall, 80 % of the studies ($N = 27$) were conducted in FL contexts and the remaining 20 % ($N = 6$) in SL contexts where the target language is dominant.

Sample sizes ranged from 16 to 144, with an average of 47 (SD = 30) and a median of 35 participants. In total, the samples amounted to 1,538 participants, which in most studies (78 %) were university students with ages ranging from 18 to 50. Three studies were conducted at language institutes, three in high schools, and one in middle school, and these studies reported lower age ranges (12–18). Proficiency levels varied across studies. More than 50% of the studies reported intermediate proficiency. A few studies examined beginners (Bueno Alastuey & Martínez de Lizarrondo Larumbe, 2017; Oh, 2014; Tian, 2011), advanced (Alshalan, 2016, Strobl, 2014 and Wigglesworth & Storch, 2009), and mixed proficiency levels (McDonough et al., 2018; Pae, 2011; Stell, 2018). Five studies did not report proficiency information. However, it is worth noting that these levels may not reference the exact proficiency level in all 33 studies, especially because proficiency measures substantially varied across studies. In 15 studies, institutional status was used as a proxy for L2 proficiency. In-house assessments which involved university entrance exams and program diagnostic tests were used in 10 studies. For these two measures, researchers often referred to the Common European Framework of Reference (CEFR), Test of English as a Foreign Language (TOEFL), or the International English Language Testing System (IELTs) levels as a means of describing learners' proficiency, yet five studies reported institutional levels with no reference to frameworks that would allow readers to interpret the proficiency levels. In seven studies, researchers used the Oxford Placement Test (OPT) and Institutional TOEFL (ITP) as standardized tests to assess proficiency. Only one study in the database characterized the participants' proficiency impressionistically and referenced the CEFR descriptors to corroborate their proficiency evaluation.

Regarding task type and mode, the database included 18 studies in which students collaborated to generate content in response to written prompts (i.e., content-generating tasks), whereas the remaining 15 studies required the students to reproduce content presented to them using pictures or graphs ($N = 6$), reading passages ($N = 6$), or oral passages ($N = 3$). Most of the collaborative tasks were completed face to face (76 %), with only seven computer-mediated studies and one study incorporating both modes.

4.2. Definition and measurement of CW outcomes

The first research question asked how the language learning outcomes of CW have been defined and measured. Table 2 summarizes the designs, dependent variables, and measures implemented across the 33 studies, 18 in one-shot designs (12 between-groups and 6 within-groups) and 15 in experimental designs. The choice of research design resulted in a relative preference for different dependent variables. In this section I discuss the dependent variables employed in the studies in some detail.

CAF measures featured centrally as dependent variables in one-shot design studies. Syntactic complexity was examined in 14 studies, all but one featuring one-shot designs. Eight different complexity measures were used. Most common were subordination

Table 2
Research Design, Dependent Variables, and Measures.

		One-Shot Design ($N = 18$)	Experimental Design ($N = 15$)
Research Design	Between-groups	$k = 12$	$k = 15$
	Within-groups	$k = 6$	NA
Dependent Variables	CAF	$k = 8$	$k = 1$
	Rubric Scores	$k = 2$	$k = 8$
	CAF + Rubric Scores	$k = 8$	$k = 1$
	Grammar/vocabulary Test Scores	NA	$k = 4$
	Rubric + Grammar/vocabulary Test Scores	NA	$k = 1$
Accuracy Measures	Ratio Measures	EFC/C (8); EFT/T (8); E/W (7); E/T (2)	EFC/C (1); EFT/T (1)
	Weighted-Ratio Measures	NA	WCR (1)
	Frequency Measures	Empty Categories (1)	NA
Complexity Measures	Subordination	C/T (9); DC/C (5); DC/T (1); C/W (1); DC/IC (1)	C/T (1)
	Global Complexity	W/T (4)	W/T (1)
	Phrasal Complexity	W/C (2)	NA
	Coordination	T/S (1)	NA
Fluency Measures	Frequency Measures	W (14); T (6); C (6); S (1); WT (1)	NA
	Ratio Measures	W/C (2); W/T (2); W/EFT (1); Ch/S (1); Ch/M (1)	NA
Rubric Scores	Macro-criteria	$k = 5$	$k = 1$
	Macro & micro-criteria	$k = 5$	$k = 9$
Test scores	Grammar test	NA	$k = 3$
	Vocabulary test	NA	$k = 2$

E = error; C = clause; Ch = characters; D = dependent; F = free; I = independent; S = sentence; T = T-unit; W = word; WT = word types; WCR = weighted clause ratio.

Table 3
Overall Mean Effect Sizes for One-Shot Design Studies (N = 18).

	K	*Hedges'g*	*SD*	95% *CIs*	
				Lower	Upper
Complexity	13	−0.14	0.21	−0.55	0.27
Accuracy	13	0.73	0.11	0.51	0.94
Fluency	15	0.10	0.19	−0.28	0.48
Rubric scores	10	0.35	0.16	0.04	0.65

measures, of which one or more were used in 12 studies. Length-based measures, argued by Norris and Ortega (2009) to gauge global and phrasal complexity, were used in seven studies altogether, usually in conjunction with subordination. Coordination complexity was only examined in one study. Accuracy was examined in 15 studies, the majority ($N = 13$) following one-shot designs. Six different accuracy measures were used. Most common were ratio indices that measure the distribution of errors in the text. One study employed a weighted clause ratio (WCR) to account for error gravity, and another study measured accuracy in Chinese as the frequency of empty categories (e.g., subject) accurately omitted in salient contexts. Most studies included two or three measures of accuracy, and only four studies used a single accuracy measure. Fluency was examined in 15 studies, all employing one-shot designs. Ten different fluency measures were used, the majority measuring amount of production in frequencies. Less common were ratio measures that measure length of production units, and all of these were used in conjunction with frequency measures. Only one study operationalized fluency as writing speed (i.e., number of characters per minute).

Rubric scores were used as a dependent variable in an equal number of one-shot ($N = 10$) and experimental studies ($N = 10$). The majority of these studies ($N = 18$) utilized analytic rubrics, except for two one-shot design studies which used holistic rubrics. The rubric criteria varied across studies: Studies that used rubric scores as the main dependent variable (2 out of 10 one-shot design; 9 out of 10 experimental design) tended to utilize rubrics that emphasized micro-criteria (e.g., grammar and vocabulary) in addition to macro-criteria (e.g., organization and structure). Conversely, studies that also measured accuracy and complexity (8 out of 10 one-shot design; 1 out of 10 experimental design) tended to weigh more on macro- rather than micro-criteria, since the latter are presumably captured by CAF measures.

Finally, grammar/vocabulary tests were used in experimental studies only ($N = 5$). Grammar posttests were used in three studies including error-correction (Alammar, 2017), passive structure recognition (Kuiken & Vedder, 2002), and grammaticality judgment

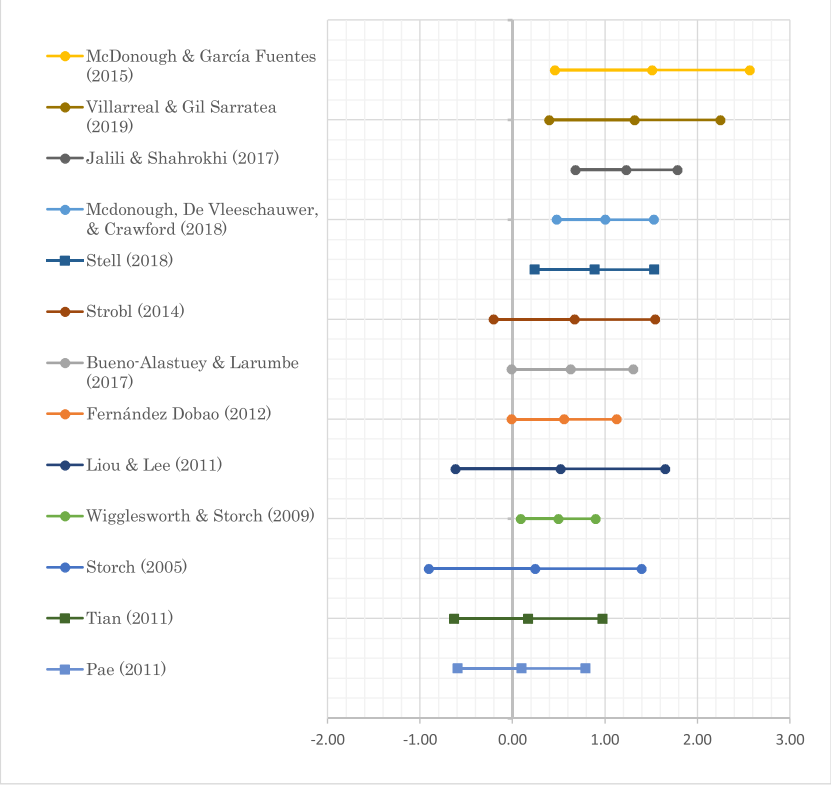

Fig. 2. Accuracy Effect Sizes & 95 % CIs Across One-Shot Design Studies (N = 13).

posttests (Reinders, 2009). The remaining two studies tested receptive knowledge of vocabulary (Liu & Lan, 2016) or required writers to compose sentences with target vocabulary (Kim, 2008).

The above characterization of designs, dependent variables, and measures, in response to Research Question 1, has built an interpretive context for the meta-analytic results of Research Question 2, which asked what the average difference is between texts written collaboratively and individually across one-shot design studies, and Research Question 3, which asked what the average difference is between posttest individual performance between CW and individual writing groups.

4.3. One-shot design: Evidence from accumulated studies

Table 3 summarizes the average effect sizes, SDs, and CIs for each of the dependent variables examined in one-shot design studies. Note that the same studies contributed to more than one dependent variable, especially in the case of CAF, which were examined in tandem in 13 of the 18 studies. The results in Table 3 are presented with reference to *Hedges'g* because this effect size estimate can be more sensitive than *Cohen's d* in small samples (Boreinstein et al., 2011).

The results from the 13 studies that compared the accuracy of collaborative and individual texts suggest that collaborative texts are more accurate than individual texts by more than two thirds SD units ($g = 0.73$) The difference can be considered statistically trustworthy, since the CIs are relatively narrow (plus or minus 0.21 SD units) and their lower boundary differed positively from zero by 0.51 SD units. The small standard deviation $SD = 0.11$ suggests that the effect sizes from the 13 studies are clustered around the mean. Based on the SLA-specific benchmarks proposed by Plonsky and Oswald (2014), CW seems to have a medium positive effect on text accuracy.

Unlike accuracy, the accumulated one-shot design studies do not yield statistically trustworthy results for the mean effect sizes of complexity, fluency, or rubric scores (see Table 3). The SDs associated with these variables are substantially large relative to corresponding mean effect sizes, which suggests that individual effect sizes are widely dispersed around the means. Further, the CIs are extremely wide for all three mean effect sizes (plus or minus 0.42 SD units for complexity; plus or minus 0.38 SD units for fluency; plus or minus 0.30 SD units for rubric scores), with lower boundaries that are either estimated at or cross zero.

The average effect sizes for the four constructs in Table 3 are better put into perspective by inspecting the individual effect sizes contributing to the aggregate effects. In what follows, therefore, I examine in turn accuracy, which yielded a positive if medium advantage for CW, and complexity, fluency, and rubric scores, neither of which showed trustworthy effects.

Fig. 2 shows a forest plot of the individual 13 studies contributing to the mean positive effect for accuracy ($g = 0.73$), each with

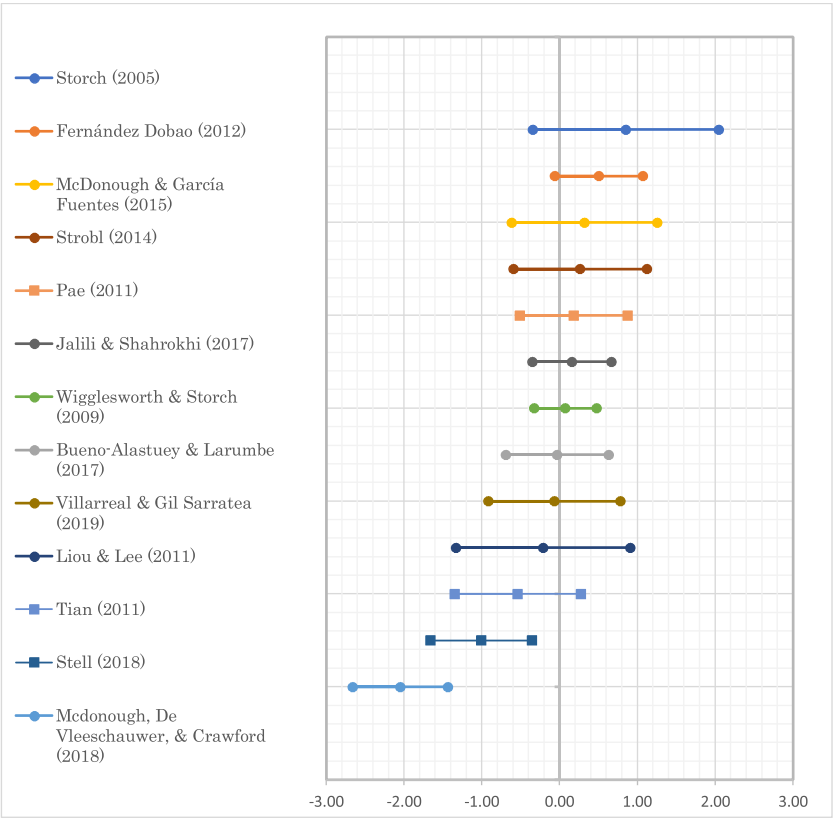

Fig. 3. Complexity Effect Sizes & 95 % CIs Across One-Shot Design Studies (N = 13).

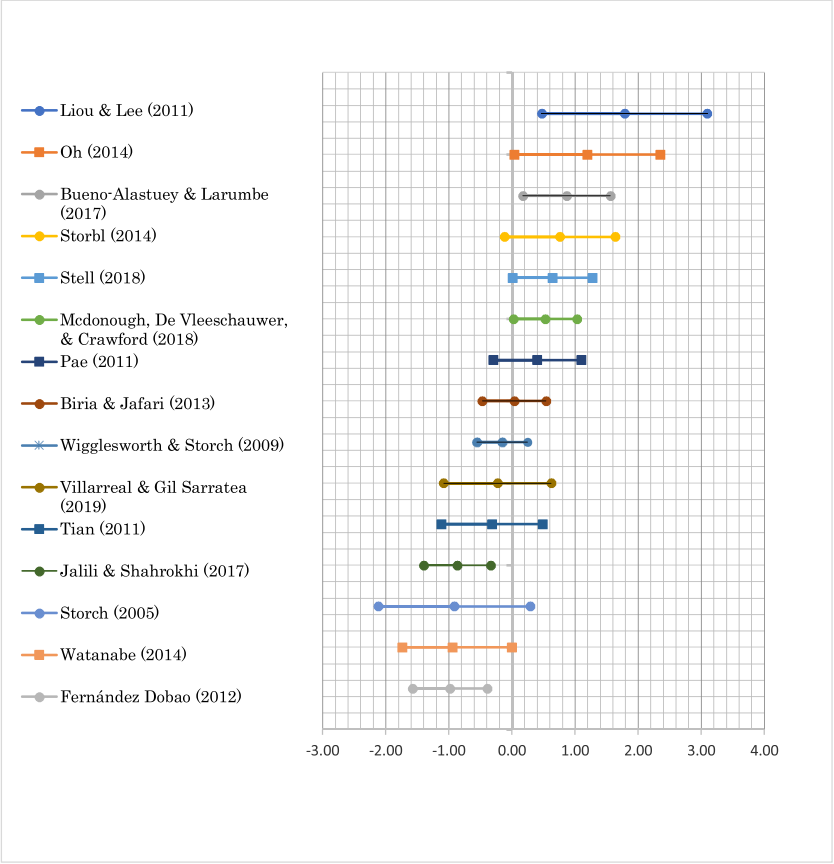

Fig. 4. Fluency Effect Sizes & 95 % CIs Across One-Shot Design Studies (N = 15).

their accuracy effect sizes and 95 % CIs. A different color marks each individual study, and the different shape bars (i.e., a filled circle vs. filled square) represent the two research approaches possible in one-shot designs that inspected accuracy: between-groups ($N = 10$) and within-groups ($N = 3$). Fig. 2 illustrates that while the magnitude of effect sizes varied across studies, each individual study yielded an observed mean effect that fell on the positive side of the magnitude scale. However, studies in which the number of collaborative or individual texts is fewer than ten are associated with wider standard error as suggested by their wide CIs. Further, seven studies have CIs crossing or very close to zero, indicating that the observed effects in these studies may not differ probabilistically from the null hypothesis. A closer analysis revealed that six out of these seven studies involved a content-reproducing task. In contrast, all studies resulting in statistically trustworthy effects for accuracy (i.e., with CIs that fall above zero) featured content-generating tasks. Thus, if more studies accumulate in the future, a moderator worthy of consideration may be whether the benefits of CW are inspected in content-generating vs. content-reproducing writing tasks.

In contrast to the favorable picture for accuracy gleaned across individual one-shot CW studies, substantial degree of heterogeneity and lack of statistical trustworthiness can be observed in the forest plots shown in Fig. 3 for complexity, Fig. 4 for fluency, and Fig. 5 for rubric scores.

As for accuracy, in Figs. 3–5 the width of the CIs for complexity, fluency, and rubric scores depended on sample size: Studies in which the number of collaborative or individual texts was under ten were associated with wider CIs and are thus less precise in their estimation of the effect size. Further, most of the CIs cross zero (i.e., 11 out of 13 for complexity; 12 out of 15 for fluency; 9 out of 10 for rubric scores). This pattern suggests that the observed effects are not probabilistically different from the null hypothesis and could be attributed to chance. This heterogeneity and the lack of statistical trustworthiness preclude a meaningful interpretation of the mean effect sizes for these three constructs in Table 3 and suggest that improved research quality across individual studies may be warranted. I will elaborate on the issue in the Discussion section.

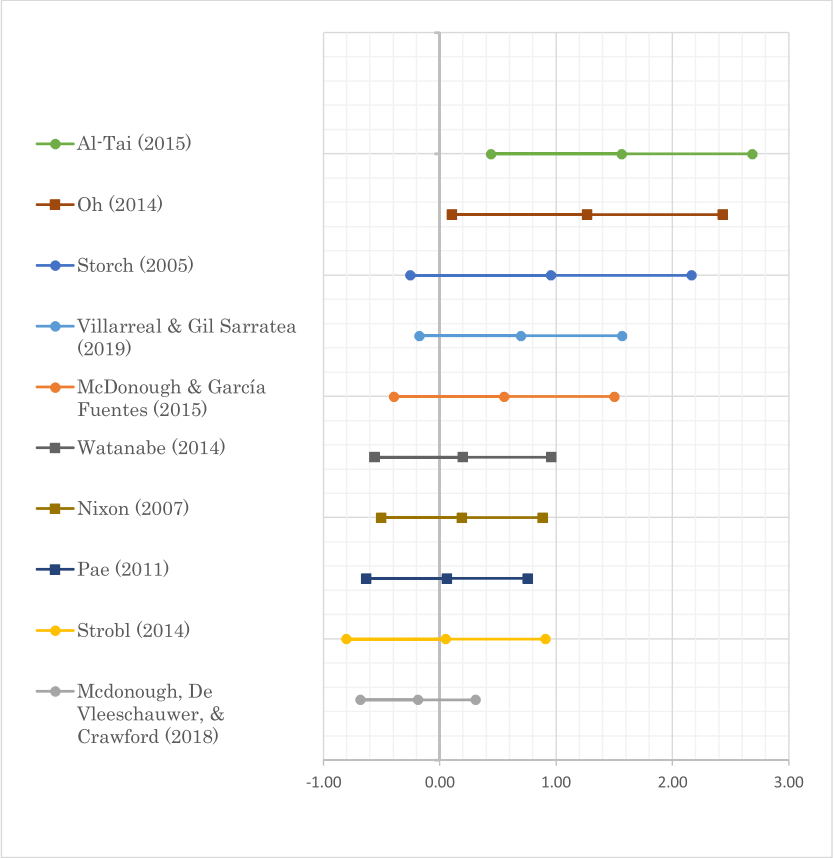

Fig. 5. Rubric Scores Effect Sizes & 95 % CIs Across One-Shot Design Studies (N = 10).

4.4. Experimental designs: evidence from accumulated studies

Table 4 summarizes the average effect sizes, SDs, and CIs for the two dependent variables examined in experimental studies: rubric scores and grammar/vocabulary test scores. (Two experimental design studies also included CAF measures, cf. Table 2, but they provided insufficient data points to aggregate into average effects.) The mean effect size from the 10 studies examining the learning gains of CW on the rubric scores of individual writing posttests indicates that students who experience CW are able to later produce individually written texts that are approximately one SD unit better ($g = 0.94$)than the texts written by others who did not experience CW. The relatively small SD (0.16)suggests that the ten effect sizes are tightly clustered around the mean. By necessity of the small sample ($k = 10$), the CIs are somewhat wide (plus or minus 0.32 SD units). Still, the lower boundary differed positively from zero by 0.62 SD units, indicating that the mean advantage can be deemed statistically trustworthy. Based on Plonsky and Oswald's (2014) benchmarks, the experience of CW seems to produce a large positive effect on subsequent text quality, as reflected in higher rubric scores on individual posttests.

Fig. 6 amplifies the finding for rubric scores across the ten available individual studies. All ten effect sizes appear on the positive side of the magnitude scale. However, the precision of observed effects differs depending on sample size. The few studies with wide CIs had less than 15 participants in one or both groups. Further, five studies have CIs crossing or very close to zero, indicating that the observed effects may not be trustworthy. Nevertheless, given the overall homogenous distribution of effect sizes (SD = 0.16) in Fig. 6 and the fact that the aggregate effect size has been weighted by the inverse of variance of these individual effects, the large average magnitude of g = 0.94 in Table 4 can be considered statistically trustworthy.

Unlike rubric scores, the average effect size for grammar/vocabulary posttest scores shown in Table 4 ($g = 0.62$), was not statistically trustworthy (95% CIs $= -0.01 - 1.22$). The SD of 0.31is substantially large and the CIs are extremely wide (plus or minus 0.61 SD units), with lower boundaries estimated at zero. The substantial degree of heterogeneity and lack of statistical trustworthiness can be observed at the individual study level in Fig. 7. While all five studies yielded positive effect sizes, the lower bands of the CIs cross zero in four of them, indicating that the observed effects may not be trustworthy.

Table 4
Overall Mean Effect Sizes for Experimental Design Studies.

	N	*Hedges'g*	*SD*	95% *CIs*	
				Lower	Upper
Rubric scores	10	0.94	0.16	0.62	1.26
Grammar/vocabulary posttests	5	0.62	0.31	0.01	1.22

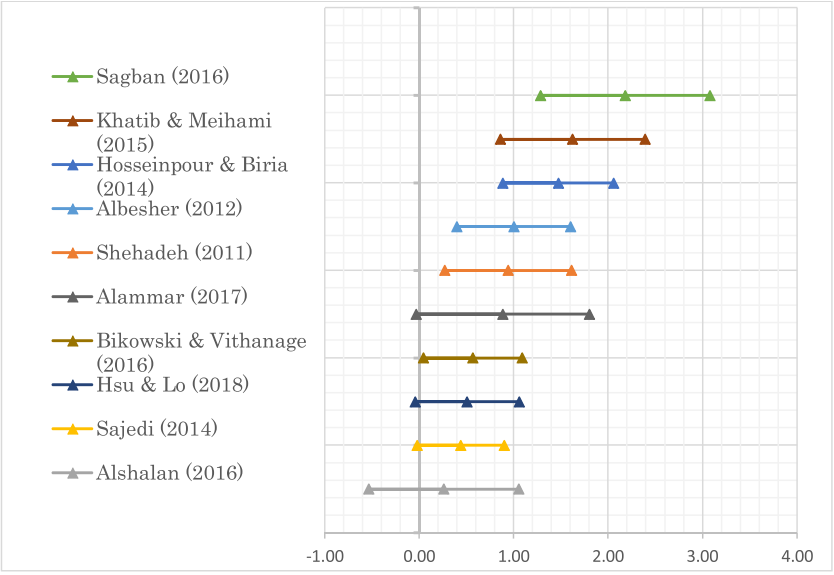

Fig. 6. Rubric Scores Effect Sizes & 95 % CIs Across Experimental Design Studies (N = 10).

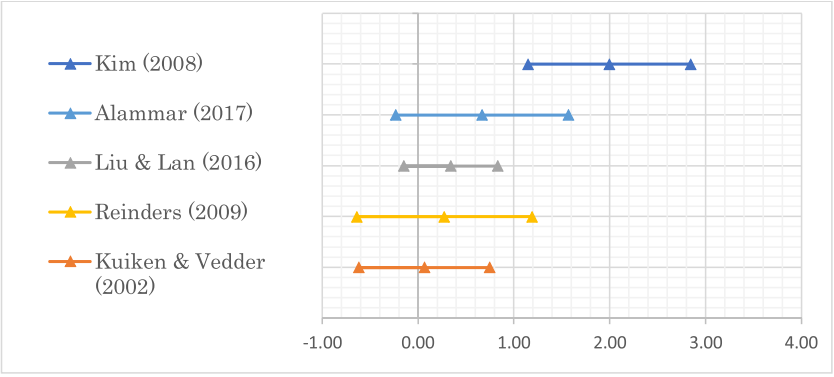

Fig. 7. Test Scores Effect Sizes & 95 % CIs Across Experimental Design Studies (N = 5).

5. Discussion

The current meta-analysis examined the accumulated empirical evidence for the claim that CW is beneficial for L2 learning and writing. The systematic synthesis of 33 studies revealed that the benefits of CW have been equally examined using two research designs. In one-shot designs, researchers have examined CAF-related and (less often) rubric score differences between texts written collaboratively versus individually within a single session, often by different writers, sometimes by the same writers. In experimental designs, researchers have examined the effects of CW on the quality of subsequent individual writing via rubric scores or (less often) grammar/vocabulary test scores. Collaboratively written texts were found to be more accurate than individually written texts by two thirds of a SD unit $g = 0.73$ ($SD = 0.11$), an effect considered of a medium magnitude (Plonsky & Oswald, 2014). Writers who experienced CW (in an experimental condition) were also found to subsequently score higher than writers who only experienced individual writing (in a control condition) when both groups' individual posttest writing was assessed through rubrics. The difference

approximated one SD unit, $g = 0.94$ ($SD = 0.16$), which can be considered large (Plonsky & Oswald, 2014). These results are indeed encouraging when compared to similar meta-analytic outcomes in the realm of L2 writing. For instance, previous meta-analyses have estimated the overall effect of written corrective feedback on L2 writing accuracy to range from small ($d = 0.14$ in Truscott, 2007) to moderate ($g = 0.54$ in Kang & Han, 2015). Therefore, the findings from the current meta-analysis supports theoretical rationales of CW as a task that enables learners to write more accurately because they can negotiate meaning (Gass & Mackay, 2007; Long, 1996), pool expertise, and engage in collective scaffolding (Donato, 1994; Lantolf & Thorne, 2007) as well as arguments for CW as a task which furnishes learners with cumulative experiences that might make them better individual writers in the long run (Storch, 2013).

For all other dependent variables submitted to the meta-analysis—complexity, fluency, and grammar/vocabulary scores—the findings were inconclusive either because accumulation was insufficient, as was the case with grammar/vocabulary posttests, or because of excessive variability across study findings, as was the case for complexity and fluency. The question is, then, what causes such variability. I argue below that much of the observed variability across findings stems from methodological considerations of two kinds: shortcomings in the use of CAF measurement, and methodological issues in one-shot designs. In addition, I suggest some systematic variability may be at work that will need to be examined via moderator analyses once more product-oriented CW research accumulates.

The measurement patterns for CAF uncovered in the present study largely echo the observations made by Zhang and Plonsky (2020) for the broader CW domain they examined, inclusive of product and process studies. The large variance in CAF measurement noted in their review was also observed in the specific domain of product-oriented CW studies in this meta-analysis (Table 2). I offer several additional insights here with regard to methodological rigor in the measurement practices for CAF seen in this domain, particularly with regard to complexity and fluency.

Congruent with the findings by Zhang and Plonsky (2020), syntactic complexity was predominately measured via subordination indices. A closer look suggests that the variability in the present findings ($g = -0.15$, $SD = 0.21$) may be due in part to a mismatch between the chosen measurement indices and the participants' proficiency. Namely, no benefits for CW were found when studies employed subordination indices as the sole measure to examine beginner (e.g., Bueno Alastuey & Martínez de Lizarrondo Larumbe, 2017) or advanced texts (e.g., Wigglesworth & Storch, 2009). It is possible that subordination indices are best employed with intermediate levels of proficiency and that they are not sensitive enough to capture the effects of collaboration on text complexity for beginners, who might not be developmentally ready for subordination (as Bardovi-Harlig, 1992, claimed) or for advanced learners, who may have reached maturity in their subordination complexity (as Norris & Ortega, 2009, claimed). Therefore, it would be premature to conclude that collaboration does not influence complexity until CW studies accumulate that measure multiple complexity dimensions (e.g., global complexity, phrasal complexity, subordination, and coordination) appropriately matched to the proficiencies of the samples.

For fluency, a different methodological shortcoming may help explain the substantial heterogeneity observed in the present results. In the CAF tradition, fluency in writing has been mainly studied through frequencies (e.g., total number of words, clauses, T-units) in the language produced (Zhang & Plonsky, 2020). However, these measures are only assumed valid if task time is constant for both conditions. Yet, many studies in this meta-analysis assigned 20-60 minutes more to the CW condition than the individual writing condition (e.g., Jalili & Shahrokhi, 2017; Stell, 2018; Wigglesworth & Storch, 2009). In such cases, and since the studies do not report standardizing language produced by time-on-task, using frequency measures may bias the fluency results in favor of CW. The few studies that used fluency ratios often employed length measures that are theoretically debated in terms of whether they are best thought of as tapping into fluency (Wolfe-Quintero, Inagaki, & Kim, 1998) or complexity (Norris & Ortega, 2009). These issues complicate interpretation of any observed effect sizes for fluency. More sophisticated measures that record writers' pauses and keystroke loggings (e.g., Chukharev-Hudilainen, Saricaoglu, Torrance, & Feng, 2019; Révész, Michel, & Lee, 2019) might illuminate our understanding of the effects of CW on fluency.

Further, one-shot design studies carry their own set of methodological problems that then dampened the cumulative observations possible in the present meta-analysis. As mentioned, the effect sizes of one-shot designs were seen to be less precise when the number of texts fell below 10 for either group, which was the case for eight out of the 18 studies. In addition to the use of small samples, 10 of the 12 studies investigating between–group differences in one shot-designs featured unbalanced samples, meaning that they included a different number of observations in each group, with differences that were up to 14 observations (e.g., Jalili & Shahrokhi, 2017). Such unbalanced designs have low statistical power and may be limited in the insight they can add to the research domain (Brysbaert, 2019). To be sure, the problems of small sample size and unbalanced samples were also seen in experimental studies. Specifically, experimental studies with samples of less than 15 participants in either group showed less precise effects, and many samples were not entirely balanced in them. However, the imbalances were particularly severe in one-shot designs, where it seems many researchers took care of balancing the number of participants in the collaborative and individual groups, when in retrospect they should have balanced the number of collaborative and individual texts, since texts rather than writers constitute the unit of observation and analysis in one-shot designs. Finally, 17 of 18 one-shot design studies did not pretest writers in the collaborative and individual writing conditions prior to the writing task, thus failing to ensure an equivalent baseline performance, which diminished the trustworthiness of the results. In sum, differences in magnitude could have been clearer, in all studies but especially in those featuring one-shot designs, had researchers employed larger samples with balanced designs and comparable pre-tested baselines. But even if these methodological shortcomings are addressed, I would like to suggest that in the future CW researchers abandon one-shot designs and concentrate their efforts on experimental designs. Any difference in magnitude in one-shot design studies, even if found, is simply a difference in textual output; it gives no indication as to whether CW has supported L2 learning or writing within the session, or if the observed benefit will endure in subsequent individual performance. Such evidence can only be gleaned from experimental design studies.

The present discussion should not be taken as an argument that variability in CW effects is only an artifact of methodological

considerations. L2 writing scholars (e.g., Manchón, 2011; Storch, 2019; see also Zhang & Plonsky, 2020) have maintained that the nature and magnitude of CW outcomes are mediated by factors such as L2 proficiency (e.g., Storch & Aldosari, 2013), task type (e.g., Storch & Wigglesworth, 2007), task mode (e.g., Wang, 2015) and number of collaborators (e.g., Bueno Alastuey & Martínez de Lizarrondo Larumbe, 2017). While a statistical moderator analysis was not possible in the present meta-analysis due to insufficient research accumulation, it will be remembered that studies that employed content-generating tasks tended to show a more statistically trustworthy positive effect on accuracy than content-reproducing tasks. Thus, task type tentatively emerges as a potential moderator of the effects of CW on accuracy worthy of future research. The need to convey predetermined content in content-reproducing tasks may have possibly taken precedence over accuracy during the negotiations and also required learners to use vocabulary and grammatical structures that are beyond their levels, which could have rendered collaborative deliberations over accuracy unsuccessful. In contrast, the content in content-generating tasks is left open for negotiation, which may allow learners to negotiate structures within their interlanguage systems and result in clearer gains in accuracy. These observations, as already cautioned, are only tentative. It is crucial to evaluate statistical and methodological rigor of any future research base for the study of the benefits of CW for L2 learning and writing, and to let the research base accumulate, before moderators can be systematically addressed.

6. Limitations and conclusion

This meta-analysis synthesized 33 product-oriented studies of L2 writing designed to explore the effects of CW on language learning. The aggregated results suggested that compared to individually written texts, collaboratively written texts are more accurate. Further, texts written individually after the experience of CW score higher on rubric criteria, indicating a higher writing quality. However, certain design features may have diminished the statistical trustworthiness of other study findings and little could be said about purported benefits in other areas, such as complexity, fluency, or grammar/vocabulary learning. Nor could the present meta-analysis address moderator variables that have been suggested to modulate the benefits of CW for L2 learners, due to the insufficient accumulation of product oriented L2 CW studies to date. Moreover, the meta-analysis included computer-mediated CW but has less to say about the impact of this mode on CW benefits, given that most CW studies available were conducted in face-to-face environments. Likewise, the present findings may not be generalizable to all kinds of writers, since knowledge of product-oriented CW in the present meta-analysis was biased in favor of EFL intermediate-level university students (as is knowledge of CW in general, see Zhang & Plonsky, 2020).

Despite these limitations, this meta-analysis makes methodological and pedagogical contributions. In terms of methodology, some recommendations can be distilled from the meta-analysis that can help advance this research domain. First, future studies should maximize the statistical trustworthiness of the findings by establishing comparable baselines via pretests, recruiting more participants, and employing balanced designs. Future studies should also resist claiming language or writing development based on designs that only compare collaborative and individual texts within a single session. Claims about the learning benefits of CW are only warranted in experimental designs which incorporate individual pre- and posttests to gauge the effects of experiencing CW on subsequent individual writing. Further, when inspecting the evidence in terms of CAF benefits, CW research would profit from future studies that analyze multiple complexity dimensions appropriately matched to the proficiencies of the sample, as well as more innovative approaches to the measurement of writing fluency. Other ways to conceptualize and measure writing development at the discourse level such as genre-based analyses (e.g., Rose, 2011) can also shed light on benefits of CW for L2 writing. More research is also needed to examine the benefits of CW for L2 learning and writing by beginner and advanced students, at younger ages, and in SL contexts. Finally, in the future it may be fruitful to investigate if, as tentatively suggested by the present meta-analytic evidence, content-generating tasks might show clearer gains of CW on accuracy than content-reproducing tasks. Process-oriented studies of CW can also help systematically illuminate the qualitative differences during collaboration that may be afforded by these two task types, and which may eventually explain any differential outcomes, if observed.

In terms of pedagogical contributions, the two main conclusions of the meta-analysis will be good news for language educators: 1) Collaboratively written texts are more accurate than individually written texts. While increased text accuracy does not guarantee that learners will be able to produce more accurate texts on their own, it does suggest that learners pooled their linguistic expertise to negotiate text meaning and form. A more accurate text can also allow teachers to focus instruction and feedback on organization and structure rather than language issues. Therefore, L2 writing teachers may strategically implement collaborative writing tasks when increased text accuracy is a desired outcome. 2) The experience of CW was found to help L2 writers subsequently produce higher-scoring texts when they write alone. This finding is indeed encouraging given the pedagogical value that teachers place on rubrics in L2 writing classrooms. Therefore, incorporating collaborative writing tasks in the L2 writing classroom can improve the text quality of L2 writers. Still, practitioners are advised to consider curricular goals and desired outcomes when implementing CW tasks. In addition, the outcomes of collaborative writing tasks can hinge on the level of mutual engagement that learners exhibit during the writing process (Elabdali & Arnold, 2020). Thus, teachers are encouraged to implement pre-collaboration activities to promote engagement among group members.

Ultimately, neither the decision to incorporate CW tasks in language classrooms nor the pronouncement of empirical benefits for CW should be solely driven by language learning outcomes. Other less visible benefits, although beyond the scope of the present meta-analysis, are equally valuable. CW might afford learners opportunities to foster cultural reflection (Angelova & Zhao, 2016), learn discipline-specific content (Rice, 2007), and navigate group projects (Ede & Lunsford, 1990), all crucial skills in their future personal and professional lives. Nevertheless, the focus of the current meta-analysis has been on language learning outcomes of CW, with the hopes that the cumulative findings synthesized here will offer pedagogical insights to practitioners and future direction to L2 writing researchers.

Acknowledgements

I would like to thank Dr. Lourdes Ortega, Dr. Nicholas Subtirelu, the two anonymous reviewers, and the editor for their comments and feedback which helped me strengthen the final version of this article. I take responsibility for the shortcomings that remain.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.jslw.2020.100788.

References[2]

*Alammar, M. (2017). The role of collaborative vs. individual writing in improving essay writing: a case study on Saudi learners. *International Journal of Arts & Sciences, 10*(2), 653–667.

*Alshalan, A. M. (2016). *The effects of wiki-based collaborative writing on ESL student's individual writing performance* (Doctoral dissertation). Retrieved from ProQuest Dissertations & Theses Global. (Order No. 10153432).

Angelova, M., & Zhao, Y. (2016). Using an online collaborative project between American and Chinese students to develop ESL teaching skills, cross-cultural awareness and language skills. *Computer Assisted Language Learning, 29*(1), 167–185. https://doi.org/10.1080/09588221.2014.907320.

Arnold, N., Ducate, L., & Kost, C. (2012). Collaboration or cooperation? Analyzing group dynamics and revision process in wikis. *CALICO Journal, 29*(3), 431–448. https://www.jstor.org/stable/calicojournal.29.3.431.

Bardovi-Harlig, K. (1992). A second look at T-unit analysis: Reconsidering the sentence. *TESOL Quarterly, 26*, 390–395. https://www.jstor.org/stable/3587016.

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2011). *Introduction to meta-analysis*. Hoboken, NJ: John Wiley & Sons.

Brysbaert, M. (2019). How many participants do we have to include in properly powered experiments? A tutorial of power analysis with reference tables. *Journal of cognition, 2*(1), 1–38. DOI: 10.5334/joc.72.

*Bueno Alastuey, M. C., & Martínez de Lizarrondo Larumbe, P. (2017). Collaborative writing in the EFL Secondary Education classroom: Comparing triad, pair and individual work. *Huarte De San Juan. Filología y Didáctica de la Lengua, 17*, 254–275. https://hdl.handle.net/2454/28503.

Chukharev-Hudilainen, E., Saricaoglu, A., Torrance, M., & Feng, H.-H. (2019). Combined deployable keystroke logging and eye tracking for investigating L2 writing fluency. *Studies in Second Language Acquisition, 41*(3), 583–604. DOI: https://doi.org/10.1017/S027226311900007X.

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). Norwood, NJ: Ablex.

Durlak, J. A. (2009). How to select, calculate, and interpret effect sizes. *Journal of pediatric psychology, 34*(9), 917–928. https://doi.org/10.1093/jpepsy/jsp004.

Ede, L., & Lunsford, A. (1990). *Singular texts/plural authors*. Carbondale: Southern Illinois University Press.

Elabdali, R., & Arnold, N. (2020). Group dynamics across interaction modes in L2 collaborative wiki writing. *Computers & Composition., 58*, 1–17. https://doi.org/10.1016/j.compcom.2020.102607.

Elola, I., & Oskoz, A. (2010). Collaborative writing: Fostering foreign language and writing conventions development. *Language Learning and Technology, 14*(3), 51–71. http://llt.msu.edu/vol14num3/elolaoskoz.pdf.

*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.

Gass, S. M., & Mackay, A. (2007). Input, interaction and output in second language acquisition. In B. van Patten, & J. Williams (Eds.), *Theories in second language acquisition: An introduction* (pp. 175–198). Mahwah, NJ: Lawrence Erlbaum Associates.

*Hsu, H., & Lo, Y. (2018). Using wiki-mediated collaboration to foster L2 writing performance. *Language Learning & Technology, 22*(3), 103–123. DOI: https://doi.org/10125/44659.

*Jalili, M. H., & Shahrokhi, M. (2017). Impact of collaborative writing on the complexity, accuracy, and fluency of Iranian EFL learners' L2 writing. *Journal of Applied Linguistics and Language Research, 4*(4), 13–28.

Kang, E., & Han, Z. (2015). The efficacy of written corrective feedback in improving L2 written accuracy: A meta-analysis. *The Modern Language Journal, 99*(1), 1–18. https://doi.org/10.1111/modl.12189.

*Khatib, M., & Meihami, H. (2015). Languaging and writing skill: The effect of collaborative writing on EFL students' writing performance. *Advances in Language and Literary Studies, 6*(1), 203–211. https://doi.org/10.7575/aiac.alls.v.6n.1p.203.

*Kim, Y. (2008). The contribution of collaborative and individual tasks to the acquisition of L2 vocabulary. *Modern Language Journal, 92*(1), 114–130. https://doi.org/10.1111/j.1540-4781.2008.00690.x.

*Kuiken, F., & Vedder, I. (2002). The effect of interaction in acquiring the grammar of a second language. *International Journal of Educational Research, 37*(3–4), 343–358. https://doi.org/10.1016/S0883-0355(03)00009-0.

Lantolf, J., & Thorne, S. (2007). Sociocultural theory and second language learning. In B. van Patten, & J. Williams (Eds.), *Theories in second language acquisition: An introduction* (pp. 201–224). Mahwah, NJ: Lawrence Erlbaum Associates.

Lee, J., Jang, J., & Plonsky, L. (2014). The effectiveness of Second Language pronunciation instruction: A meta-analysis. *Applied Linguistics, 36*(3), 345–366. https://doi.org/10.1093/applin/amu040.

Li, M. (2018). Computer-mediated collaborative writing in L2 contexts: An analysis of empirical research. *Computer Assisted Language Learning, 31*(8), 882–904. https://doi.org/10.1080/09588221.2018.1465981.

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

*Liou, H., & Lee, S. (2011). How wiki-based writing influences college students' collaborative and individual composing products, processes, and learners' perceptions. *International Journal of Computer Assisted Language Learning and Teaching, 1*(1), 45–61. https://doi.org/10.4018/ijcallt.2011010104.

*Liu, S. H.-J., & Lan, Y.-J. (2016). Social constructivist approach to web-based EFL learning: Collaboration, motivation, and perception on the use of Google Docs. *Educational Technology & Society, 19*(1), 171–186. https://www.jstor.org/stable/jeductechsoci.19.1.171.

Long, M. (1996). The role of the linguistic environment in second language acquisition. In W. Ritchie, & T. Bhatia (Eds.), *Handbook of second language acquisition* (pp. 413–468). San Diego: Academic Press.

Manchón, R. (2011). Writing to learn the language. Issues in theory and research. In R. Manchón (Ed.), *Learning-to-Write and Writing-to-Learn in an Additional Language* (pp. 61–82). Amsterdam: Benjamins Pub. Co.

[2] Note: Asterisked studies were included in the meta-analysis.

McAndrews, M. (2019). Short periods of instruction improve learners' phonological categories for L2 suprasegmental features. *System, 82*, 151–160. https://doi.org/10.1016/j.system.2019.04.007.
*McDonough, K., De Vleeschauwer, J., & Crawford, W. (2018). Comparing the quality of collaborative writing, collaborative prewriting, and individual texts in a Thai EFL context. *System, 74*(1), 109–120. https://doi.org/10.1016/j.system.2018.02.010.
Norris, J. M., & Ortega, L. (2000). Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis. *Language learning, 50*(3), 417–528. https://doi.org/10.1111/0023-8333.00136.
Norris, J. M., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied linguistics, 30*(4), 555–578. https://doi.org/10.1093/applin/amp044.
*Oh, H. (2014). Learners' writing performance, revision behavior, writing strategy, and perception in wiki-mediated collaborative writing. *Multimedia-Assisted Language Learning, 17*(2), 176–199. Retrieved from http://journal.kamall.or.kr/wp-content/uploads/2014/07/Oh_17_2_08.pdf.
Oswald, F. L., & Plonsky, L. (2010). Meta-analysis in second language research: Choices and challenges. *Annual Review of Applied Linguistics, 30*, 85–110. https://doi.org/10.1017/S0267190510000115.
*Pae, J. (2011). Collaborative writing versus individual writing: Fluency, accuracy, complexity, and essay score. *Multimedia-Assisted Language Learning, 14*(1), 121–148. Retrieved from http://kamall.or.kr/kor/publictions/MALL/14-1-2011.pdf#page=121.
Plonsky, L., & Zhuang, J. (2019). A meta-analysis of L2 pragmatics instruction. In N. Taguchi (Ed.), *Routledge handbook of second language acquisition and pragmatics*. New York, NY: Routledge.
Plonsky, L., & Brown, D. (2015). Domain definition and search techniques in meta-analyses of L2 research (Or why 18 meta-analyses of feedback have different results). *Second Language Research, 31*(2), 267–278.
Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning, 64*(4), 878–912. https://doi.org/10.1177/0267658314536436.
*Reinders, H. (2009). Learner uptake and acquisition in three grammar-oriented production activities. *Language Teaching Research, 13*(2), 201–222. https://doi.org/10.1177/1362168809103449.
Révész, A., Michel, M., & Lee, M. (2019). Exploring second language writers' pausing and revision behaviors: A mixed-methods study. *Studies in Second Language Acquisition, 41*(3), 605–631. https://doi.org/10.1017/S027226311900024X.
Rice, J. A. (2007). Bridging the gap: Contextualizing professional ethics in collaborative writing projects. *Business Communication Quarterly, 70*(4), 470–475. https://doi.org/10.1177/1080569907070040306.
Rose, D. (2011). Genre in the Sydney school. In J. Gee, & M. Handford (Eds.), *The Routledge handbook of discourse analysis* (pp. 209–225). London: Routledge.
*Shehadeh, A. (2011). Effects and student perceptions of collaborative writing in L2. *Journal of Second Language Writing, 20*(4), 286–305. https://doi.org/10.1016/j.jslw.2011.05.010.
Skehan, P. (2003). Task-based instruction. *Language teaching, 36*(1), 1–14. https://doi.org/10.1017/S026144480200188X.
*Stell, A. (2018). Exploring the use of collaborative writing in an EFL classroom context. *University of Sydney Papers in TESOL, 13*, 63–97. Retrieved from https://faculty.edfac.usyd.edu.au/projects/usp_in_tesol/pdf/volume13/Article03.pdf.
*Storch, N. (2005). Collaborative writing: Product, process, and students' reflections. *Journal of Second Language Writing, 14*(1), 153–173. https://doi.org/10.1016/j.jslw.2005.05.002.
Storch, N. (2013). *Collaborative writing in L2 classrooms*. Bristol, UK: Multilingual Matters.
Storch, N. (2018). Collaborative writing. *The TESOL Encyclopedia of English Language Teaching* (pp. 1–6). https://doi.org/10.1002/9781118784235.eelt0395.
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: Comparing individual and collaborative writing. In M. P. Garcia Mayo (Ed.), *Investigating tasks in formal language learning* (pp. 157–177). Buffalo, NY: Multilingual Matters.
*Strobl, C. (2014). Affordances of web 2.0 technologies for collaborative advanced writing in a foreign language. *CALICO Journal, 31*(1), 1–18. https://www.jstor.org/stable/calicojournal.31.1.1.
*Tian, J. (2011). *The effects of peer editing versus co-writing on writing in Chinese-as-a-foreign language* (Doctoral dissertation). Retrieved from ProQuest Dissertations & Theses Global. (Order No. NR94743).
Truscott, J. (2007). The effect of error correction on learners' ability to write accurately. *Journal of Second Language Writing, 16*, 255–272. https://doi.org/10.1016/j.jslw.2007.06.003.
van Aert, R. C., Wicherts, J. M., & van Assen, M. A. (2019). Publication bias examined in meta-analyses from psychology and medicine: A meta-meta-analysis. *PLoS One, 14*(4), 1–32. https://doi.org/10.1371/journal.pone.0215052.
Wang, Y. (2015). Promoting collaborative writing through wikis: A new approach for advancing innovative and active learning in an ESP context. *Computer Assisted Language Learning, 28*(6), 499–512. https://doi.org/10.1080/09588221.2014.881386.
*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.
Wolfe-Quintero, K., Inagaki, S., & Kim, H. Y. (1998). *Second language development in writing: Measures of fluency, accuracy, & complexity (No. 17)*. University of Hawaii Press.
*Wigglesworth, G., & Storch, N. (2009). Pair versus individual writing: Effects on fluency, complexity and accuracy. *Language Testing, 26*(3), 445–466. https://doi.org/10.1177/0265532209104670.
Zhang, M., & Plonsky, L. (2020). Collaborative writing in face-to-face settings: A substantive and methodological review. *Journal of second language writing, 49*. https://doi.org/10.1016/j.jslw.2020.100753.

Further reading

*Albesher, K. (2012). *Developing the writing skills of ESL students through the collaborative learning strategy*. Doctoral dissertation. University of Newcastle Upon Tyne.
*Altai, Y. (2015). The effect of collaboration on Omani students' writing: A comparison between individual, pair and group work. *European Scientific Journal, 1*, 154–171. http://eujournal.org/index.php/esj/article/view/5554.
*Bikowski, D., & Vithanage, R. (2016). Effects of web-based collaborative writing on individual L2 writing development. *Language Learning & Technology, 20*(1), 79–99. http://llt.msu.edu/issues/february2016/bikowskivithanage.pdf.
*Biria, R., & Jafari, S. (2013). The impact of collaborative writing on the writing fluency of Iranian EFL learners. *Journal of Language Teaching and Research, 4*(1), 164–175. https://doi.org/10.4304/jltr.4.1.164-175.
*Hosseinpour, N., & Biria, R. (2014). Improving Iranian EFL learners' writing through task-based collaboration. *Theory and Practice in Language Studies, 4*(11), 2428–2435. https://doi.org/10.4304/tpls.4.11.2428-2435.
*Jafari, N., & Ansari, D. N. (2012). The effect of collaboration on Iranian EFL learners' writing accuracy. *International Education Studies, 5*(2), 125–131.
*McDonough, K., & García Fuentes, C. (2015). Writing to learn language: The effect of writing task on Colombian EFL learners' language use. *TESL Canada Journal, 32*(2), 67–79. https://doi.org/10.18806/tesl.v32i2.1208.
*Nixon, R. M. (2007). *Collaborative and independent writing among adult Thai EFL learners: Verbal interactions, compositions, and attitudes* (Doctoral dissertation). Available from ProQuest Dissertations & Theses Global. (Order No. NR27931).
Plonsky, L. (2014). Study quality in quantitative L2 research (1990–2010): A methodological synthesis and call for reform. *Modern Language Journal, 98*(1), 450–470. https://doi.org/10.1111/j.1540-4781.2014.12058.x.

*Sagban, A. A. (2016). The effect of collaborative writing activities on Iraqi EFL college students' performance in writing composition. *Basic Education College Magazine for Educational and Humanities Sciences*, (26), 269–278.
*Sajedi, S. (2014). Collaborative summary writing and EFL students' L2 development. *Procedia - Social and Behavioral Sciences, 98*, 1650–1657. https://doi.org/10.1016/j.sbspro.2014.03.589.
*Watanabe, Y. (2014). *Collaborative and independent writing: Japanese university English learners' processes, texts and opinions*. Doctoral dissertation. University of Toronto.

Rima Elabdali holds an MA in TESOL from Portland State University and is currently a doctoral student in Applied Linguistics at Georgetown University. Her research spans second language (L2) writing, teacher training, corpus linguistics, and heritage lanaguage education.