TYPE AND AMOUNT OF INPUT-BASED PRACTICE IN CALI: THE REVELATIONS OF A TRIANGULATED RESEARCH DESIGN

Luis Cerezo, American University

Research shows that computer-generated corrective feedback can promote second language development, but there is no consensus about which type is the most effective. The scale is tipped in favor of more explicit feedback that provides metalinguistic explanations, but counterevidence indicates that minimally explicit feedback of the *right/wrong* type may promote comparable learning outcomes. Addressing these conflicting findings, the present study investigated the effects of different types and amounts of practice as variables that may moderate the effectiveness of computerized *right/wrong* feedback. Fifty-two learners of intermediate Spanish completed either 28 or 56 items of an input-based task with 2 or 4 options targeting Spanish past counterfactual conditional sentences. Quantitative results on achievement scores showed that differences in amount of practice might contribute to explaining the conflicting findings in the literature. Additionally, a qualitative analysis of participant mouse-click histories illustrated participant use of elimination strategies to redefine the 4-option tasks, while participant think-alouds revealed the increased boredom and fatigue induced by the extra amount of practice. This study thus contributes to the debate on the effects of different types of computerized feedback and the development of hybrid and online language learning programs, while underscoring the importance of triangulating data from multiple sources.

Language(s) Learned in this Study: Spanish

Keywords: Corrective Feedback, Amount of Practice, Input-Based Practice, CALI, Behavior-Tracking Technology, Language Learning Strategies

APA Citation: Cerezo, L. (2016). Type and amount of input-based practice in CALI: The revelations of a triangulated research design. *Language Learning & Technology*, *20*(1), 100–123. Retrieved from http://llt.msu.edu/issues/february2016/cerezo.pdf

Received: March 6, 2013; **Accepted:** November 29, 2013; **Published:** February 1, 2016

INTRODUCTION

Increasingly, studies in second language acquisition (SLA) have employed the computer for two main purposes: as a methodological tool—for example, to randomize participants, deliver second language (L2) input consistently, collect achievement data, or record online behavior (e.g., Sanz, Morales-Front, Zalbidea, & Zárate-Sández, 2015)—and as a pedagogical tool—to either retrieve or process input, mediate communication among learners, or provide instruction (e.g., Levy, 2009). The latter case, which is often referred to as computer-assisted language instruction (CALI), constitutes the focus of this study.

In CALI the computer acts as an e-tutor, that is, software with the ability to (a) engage learners in some sort of pedagogical practice and (b) provide corrective feedback in response to errors by the learners. Hitherto, over 20 studies have shown that computer-generated corrective feedback can help learners develop their knowledge and skill of a variety of L2 grammatical structures (see Cerezo, 2012 for a review). However, in line with the more extensive literature on human-delivered feedback (see the meta-analyses by Li, 2010; Lyster & Saito, 2010; Mackey & Goo, 2007; Russell & Spada, 2006) the jury is still out as to which type of feedback is the most effective. Out of at least 11 CALI studies (including five doctoral dissertations) that have investigated this issue empirically, seven found an edge for feedback

providing grammatical rules (Bowles, 2008; Cerezo, 2010; Lado, Bowden, Stafford, & Sanz, 2014; Lin, 2009; Nagata, 1993; also reported in Nagata & Swisher, 1995; Rosa & Leow, 2004; Sachs, 2011) while four found that less informative feedback of the *right/wrong* type may be equally effective (Camblor, 2006; Hsieh, 2008; Moreno, 2007; Sanz & Morgan-Short, 2004).

The varying effectiveness of *right/wrong* feedback in these CALI studies may possibly be due to differences in their experimental treatments, including, but not limited to, the type and amount of computerized practice provided. For example, Sanz and Morgan-Short (2004) argued that their group with *right/wrong* feedback performed comparably to that with grammatical explanations, contra Rosa and Leow (2004) (where it performed significantly worse), because their treatment included more practice items (56 instead of 28), which is crucial in less explicit learning conditions. Also, while both studies used input-based practice in the form of multiple-choice interpretation tasks, these differed, among other things, in the number of choices they provided. As Sanz and Morgan-Short (2004) put it:

> Our learners were presented with only two possible answers for each item, whereas participants in "Rosa and Leow (2004)" had to decide among four options. Consequently, they were required to eliminate three incorrect options in order to identify the correct choice. [W]ithout any explicit information, 18 practice items containing the target form [plus 10 distractors, totaling 28 items] supplied along with three incorrect options simply may not have been enough evidence to be conducive for noticing or understanding the target form. (pp. 70–71)

The present study is the first to empirically investigate these claims, in an attempt to determine whether the differences in type and amount of input-based practice between Sanz and Morgan-Short (2004) and Rosa and Leow (2004) contribute to explaining the conflicting results of *right/wrong* feedback in the CALI literature. To do so, Rosa and Leow's group with *right/wrong* feedback was cloned into four experimental groups that differed in (a) the type of input-based practice (2 or 4 options) and (b) the amount of practice (28 or 56 items). Additionally, learner online behavior was recorded via mouse-click tracking and think-alouds, which served to triangulate quantitative results and revealed potentially intervening variables.

LITERATURE REVIEW

Type of Input-Based Practice

An incipient number of CALI studies (Morgan-Short & Wood Bowden, 2006; Nagata, 1998a, 1998b) suggests that output-based practice (in which learners produce the target structure) yields higher L2 grammar development than input-based practice (in which learners recognize or interpret the structure from several options) (though see the meta-analysis by Shintani, Li, & Ellis, 2013). Input-based practice, however, has been more profusely used in the literature, probably because it can be more easily programmed (the number of participant responses is limited). Specifically, input-based tasks were used in eight of the 11 (73%) CALI studies that have empirically investigated the effects of type of feedback, and in three of the four (75%) studies that found no differences between *right/wrong* feedback and feedback with grammatical explanations.

Despite this preference for input-based tasks, very few CALI studies (all unpublished doctoral dissertations) have investigated which task-features may enhance their effectiveness (see Table 1 for a summary of studies). Moreno (2007) found that pushing learners to focus their attention on the targeted structure to complete a picture interpretation task did not yield superior learning gains. Medina (2008) found that asking learners to spot and highlight the targeted structure after reading a text produced superior written production gains. And Torres (2013) found that differently complex tasks yielded similar learning gains.

Table 1. *Summary of Studies on Type of Computerized Input-Based Tasks*

Study	L2 Target Form	Participants	Task by Experimental Group	Results
Medina (2008)	Spanish imperfect subjunctive	80 learners of intermediate Spanish	[More complex]: reading a text and then re-reading it to spot and highlight the targeted structure. [Less complex]: reading a text and then re-reading it to re-highlight already-highlighted targeted forms.	The [More complex] group achieved higher accuracy in controlled written production but not recognition.
Moreno (2007)	Spanish third person pre-verbal direct object pronouns	57 learners of beginning Spanish	[+Task-essential]: selecting words in a branching tree to compose a sentence describing a picture. [-Task-essential]: matching one of two sentences to a picture (the sentences differed in constituents other than the targeted structure).	Task-essentialness did not yield any significant differences on oral picture description, written picture description, or written form recognition.
Torres (2013)	Spanish present subjunctive in adjectival relative clauses	84 learners of Spanish	[Complex]: selecting the targeted structure from two options to complete a sentence describing dorm residents doing something. Then, selecting and describing one of four pictures illustrating the reason for the resident behavior. [Noncomplex]: same as above, without having to select a picture.	Task-complexity did not yield any significant differences in oral production. However, only the [noncomplex] group outperformed the control group from pretest to delayed posttest in written production.

Given this paucity of studies, there is an urgent need to further investigate the effects of different input-based task features. According to Sanz and Morgan-Short (2004), one possible avenue is to investigate the effects of different numbers of options. While they do not provide a psycholinguistic rationale for why different numbers of options may be differently "conducive for noticing or understanding the target form" (p. 71), at least two hypotheses may be invoked. On the one hand, based on the Noticing Hypothesis (e.g., Schmidt, 2001) it could be posited that an input-based task with more options may distract learner attention away from the targeted form. On the other, based on the Depth of Processing Hypothesis (e.g., Craik & Tulving, 1975) it could be argued that the additional cognitive effort devoted to identify the targeted form from a larger number of distractors may result in more robust intake. To investigate this, the present study compared the effects of two computerized input-based tasks differing in the number of options (2 vs. 4) on L2 grammar development.

Amount of Practice

The old dictum "practice makes perfect" suggests that the amount of practice provided may moderate the effects of a task on L2 grammar development. Yet, no published CALI study has empirically investigated this. In the general SLA literature, studies can be divided into two groups. One group investigated the effects of amount of practice separately (Adams, 2003; Ahmadian, 2011; Ahmadian & Tavakoli, 2011; Bygate, 1996, 2001; Gass, Mackey, Fernández, & Álvarez-Torres, 1999; Lynch & Maclean, 2000, 2001; Wang, 2009). The other investigated the effects of amount of practice in combination with or in comparison to other variables, such as the provision of grammatical explanations, input, feedback, or planning (Adams, 2003; Ahmadian & Tavakoli, 2011; Hawkes, 2012; Leow, 1998; Sheppard, 2006; Wang, 2009). A summary of studies is included in Table 2.

The vast majority of studies in Table 2 have operationalized practice as *tasks*, investigating the developmental effects of repeating the same task instance (e.g., narrating the same video excerpt) or task type (e.g., narrating different video excerpts). Most studies have used oral tasks (e.g., conducting an interview) and minimally written tasks (e.g., narrating a picture sequence). There is great variance in the number of elicited repetitions (from 1 to 11), the intervals between them (from no break to 3 weeks), and the overall time span between pretest and posttest (from 1 minute to 6 months). Because tasks prioritize communicative goals over the usage of a specific linguistic form (e.g., Ellis, 2003), most of these studies have measured L2 development holistically, quantifying change in combined areas of language (e.g., grammar and vocabulary) in terms of complexity, accuracy, and fluency (CAF) with different units. Also, due to small sample sizes and low or inconsistent usage of linguistic forms by the learners, these studies have mostly used descriptive rather than inferential statistics.

As pointed out by several authors (e.g., Ahmadian & Tavakoli, 2011; Ellis, 2009), overall studies have shown that task repetition has clear benefits for L2 complexity and fluency, but not necessarily for accuracy—rare exceptions are the case studies by Bygate (1996) and Lynch and McLean (2000, 2001) and the experimental studies by Adams (2003), Gass et al. (1999), and Wang (2009). These results can be explained from a psycholinguistic perspective. According to Levelt's (1989) speech production model, when first completing a task learners are busy figuring out what they want to say (*conceptualization*), which leaves little attentional capacity for translating their ideas into actual speech (*encoding* and *articulation*). In subsequent repetitions, learners can concentrate on improving their language performance, but given the limited nature of attentional capacity, attending to one dimension of performance (e.g., complexity or fluency) is likely to compromise another (e.g., accuracy), as posited by the Trade-Off Hypothesis (Skehan, 2009).

Table 2. *Summary of Studies on Amount of Practice*

Study	L2 Target Form	Participants	Task and Mode	Repetitions and Time Intervals	Results (CAF)
*Adams (2003)	Spanish general vocabulary and grammar	56 learners of intermediate Spanish	Narrating a comic strip (first orally, then in writing) with or without a noticing phase, and with or without stimulated recalls	1 repetition 7 days later	Task repetition increased *A*; *A* was increased more by including a noticing phase, and even more with a stimulated recall.
Ahmadian (2011)	English general language (syntactic complexity and variety, correct clauses and verb forms, syllables per minute)	30 learners of intermediate English	Narrating a video (orally)	11 repetitions (every 2 weeks over 6 months)	Task repetition increased *C*, *F*.
*Ahmadian and Tavakoli (2011)	English general language (correct clauses and verbs forms, syntactic complexity and variety, syllables per minute)	60 learners of intermediate English	Narrating a video after watching it (orally) with or without "careful" online planning (time constraints) and with or without task repetition	1 repetition after a week	Task repetition increased *C*, *F*; careful online planning increased *C*, *A* and decreased *F*; careful online planning and task repetition increased CAF.
Bygate (1996)	English general language (strong focus on vocabulary but included grammatical markers and structure)	1 learner of English	Narrating a video (orally)	1 repetition 3 days after first task	Task repetition increased CAF.

Study	L2 Target Form	Participants	Task and Mode	Repetitions and Time Intervals	Results (CAF)
Bygate (2001)	English general language (syntax, vocabulary)	48 learners of English	(1) Narrations (2) Interviews (orally)	7 repetitions (2 repetitions every 2 weeks + posttest on week 10)	Task repetition increased *C*, *F*.
Gass et al. (1999)	Spanish overall proficiency, morphosyntax and lexical sophistication	103 learners of fourth semester Spanish (data analysis: 30 participants)	Narrating a video with the same or different content (orally)	3 repetitions (1 repetition every 2-3 days + posttest 2 weeks later)	Task repetition increased overall proficiency, lexical *C*, and morphosyntactic *A* of *estar*.
*Hawkes (2012)	English grammar and vocabulary: giving opinions, comparatives, superlatives, agreement, body parts, adjectives, future with 'going to' and 'might'	26 Japanese learners of English (5-6 dyads per task type)	Performing three oral tasks after listening to a demonstration (opinion exchange, describe and draw, timed conversation), followed by a form-focus stage	1 repetition after form-focused practice following first task	Task repetition preceded by a form focused stage increased use of targeted forms and form corrections by learners.
*Leow (1998)	Spanish stem-changing verbs in the preterit	88 learners of beginning Spanish	Completing a crossword puzzle with embedded implicit feedback (in writing)	1 repetition 3 weeks after first task	Task repetition increased *A* (form recognition and controlled production) on immediate posttest (week 3) and delayed posttest (week 14).

Study	L2 Target Form	Participants	Task and Mode	Repetitions and Time Intervals	Results (CAF)
Lynch and McLean (2000, 2001)	English general language (syntax, vocabulary, and pronunciation)	14 learners of medical English (data analysis: 4 participants)	Presenting a poster (orally)	6 repetitions with no breaks	Task repetition increased syntactic *A* and overall *F* in lower and intermediate proficiency learners.
*Sheppard (2006)	English general language	Not accessible	Completing a narrative task (orally) with or without a form-focused phase n between with either input or feedback	1 repetition after an undetermined period	Task repetition increased *C*, *F* and minimally increased *A*; receiving input increased CAF; receiving feedback increased CAF the most.
*Wang (2009)	English general language	Not accessible	Narrating a video (orally) while watching it, with or without silently pre-watching it and with either online planning (slowed-down video) or strategic planning (time to prepare)	1 repetition 1 minute after first task	Task repetition increased CAF to a greater extent than any other condition; silently pre-watching (by itself or with either strategic or online planning) increased *C*; silently pre-watching with online planning increased *A*.

*Notes. CAF = Complexity, Accuracy, Fluency. * = Studies investigating task repetition in combination with, or compared to, other pedagogical interventions.*

Granted, it could be posited that the potentially beneficial effects of task repetition on accuracy might arise in the medium or long run, rather than immediately (Larsen-Freeman, 2009). Yet, using massed repetitions over an extended period of time, Bygate (2001) and Ahmadian (2011) found positive effects for fluency and complexity only. On the other hand, task repetition might need to be supplemented with other pedagogical interventions. Several studies combined task repetition with interventions such as exposure to additional input (Sheppard, 2006); exposure to feedback, either implicitly through the task (Leow, 1998) or explicitly after task completion (Sheppard, 2006); participation in "noticing phases", in which participants compared their performance with a model (Adams, 2003); participation in form-focused activities (Hawkes, 2012); and planning, either ahead of time (Wang, 2009) or online through untimed tasks (Ahmadian & Tavakoli, 2011; Wang, 2009). Interestingly, all of these studies found significant gains in accuracy during subsequent repetitions.

Building upon this body of literature, and using *right/wrong* feedback as an additional pedagogical intervention, the present study investigated the effects of different amounts of computerized practice (28 vs. 56 items) on L2 grammar development.

RESEARCH QUESTIONS AND HYPOTHESES

Existing research presents contradicting results about the effects of different types of computerized feedback on L2 grammar development. Studies like Rosa and Leow (2004) showed an edge for feedback with grammatical explanations, while other studies like Sanz and Morgan-Short (2004) obtained that *right/wrong* feedback can promote comparable learning. Following Sanz and Morgan-Short's claims, this study is first to investigate whether type and amount of input-based practice may moderate the effects of *right/wrong* computerized feedback. Specifically, the following research questions (RQs) and hypotheses (Hs) are formulated:

RQ1. On type of input-based practice. Does type of computerized input-based practice (2 or 4 options) have differential effects on L2 development?

H1. No. Learners completing an input-based task with either 2 or 4 options will show comparable L2 development. Hitherto, no study has investigated this issue, so the null hypothesis is formulated.

RQ2. On amount of input-based practice. Does amount of computerized input-based practice (28 or 56 items) have differential effects on L2 development?

H2. Yes. Learners completing 56 items of a computerized input-based task will show significantly greater L2 development. Studies have shown that more practice combined with corrective feedback promotes higher accuracy gains.

METHOD

Participants

The participants in this study were 52 undergraduate students of intermediate Spanish at a northeastern university in the United States meeting these criteria: (a) they were native speakers of English; (b) they neither spoke nor had been formally exposed to any Romance language for more than two years; and (c) they demonstrated no ability to recognize or produce the targeted structure.

Targeted Structure

As in Rosa and Leow (2004), this study used two different types of Spanish counterfactual conditional sentences: one referring to the present or future with a result in the present (e.g., *Si Joe estudiara español, entendería el discurso de Antonio Banderas* ["If Joe studied Spanish, he would understand Antonio Banderas' speech"]), the other referring to the past with a result in the present (e.g., *Si Joe hubiera*

estudiado español, ahora entendería el discurso de Antonio Banderas ["If Joe had studied Spanish, he would now understand Antonio Banderas' speech"]). The first type (henceforth referred to as *non-targeted structure*) was used as a base structure to provide participants with a general framework on Spanish conditional sentences, and for distracting purposes. The second type was the targeted structure. Rosa and Leow selected this targeted structure for three reasons: (a) the likely unfamiliarity of intermediate Spanish students with the target rules, which minimizes a potential activation of prior knowledge; (b) the ease with which it can be contrasted with contrary-to-fact conditionals in the other tenses, present or future; and (c) the high level of complexity of the structures.

Materials

Experimental Task

The pedagogical treatment for this study replicated the input-based task used in Rosa and Leow (2004) and was administered via a computer program developed in ColdFusion (see Appendix for sample items). First, participants were presented with a context statement (e.g., *Los hermanos Wright construyeron el primer avión* ["The Wright brothers built the first airplane"]) and then they were asked to complete the condition part of a conditional sentence by selecting the right verb form from a list of options (e.g., *Hoy no podríamos viajar a Europa tan rápidamente si los Wright no* _____ *el primer avión*. ["We would not be able to travel to Europe so fast these days if the Wright brothers _____ the first airplane"]). The consequence part of the sentence was always in the simple conditional tense. Thus, to fill in the blanks participants could only rely on understanding the initial context statement, rather than language cues.

The experimental groups received different numbers of options per item (2 or 4) and practice items (28 or 56). The task for the two-option groups included (a) the targeted structure, that is, the Spanish pluperfect subjunctive (e.g., *no hubieran construido* ["hadn't built"]) and (b) the non-targeted structure, the Spanish imperfect subjunctive (e.g., *no construyeran* ["didn't build"]). The four-option groups additionally included (c) the Spanish simple conditional (e.g., *no construirían* ["wouldn't build"]) and (d) the Spanish perfect conditional (e.g., *no habrían construido* ["wouldn't have built"]), alternating every other item with an ungrammatical version of the Spanish pluperfect subjunctive (e.g., **no habueran construido*). As in Rosa and Leow (2004), only options (a) and (b) were possible answers throughout the treatment, respectively providing the right choice for the critical items—which elicited the targeted structure—and the distractors—which elicited the non-targeted structure. All options were presented in randomized order. The treatment for the 28-item group contained 18 critical items and 10 distractors, while the 56-item group completed double as many items (36 critical items and 20 distractors). All experimental groups received concurrent feedback of the *right/wrong* type ("Cool!" and "Oops, try again!").

Assessment Tasks

The assessment tasks included a 34-item multiple-choice recognition task and a 32-item controlled production task in the written mode, each with 20 critical items. The recognition task had the same structure as the experimental task, while the production task did not provide any options[1]. Pretests and posttests included the same items, in randomized order.

Procedure

On day one, all the participants completed the pretests and received basic preliminary instruction on the non-targeted structure. Participants who scored more than 3 correct items (out of 20) on the written recognition pretest and/or more than 0 correct items on the controlled written production pretest were excluded from the final pool. Two days later, the remaining 52 participants were randomly assigned to one of four treatment conditions, as per the interaction of the two experimental variables in the study: type of input-based practice (2 or 4 options) and amount of input-based practice (28 or 56 items). Henceforth, these groups will be referred to as 2x28, 4x28, 2x56, and 4x56. Before completing the treatment,

participants were instructed to "think aloud" (i.e., verbalize whatever crossed their minds) throughout the entire experiment. To familiarize participants with the procedure, they received this instruction while performing a mathematical problem-solving task. Next, they completed the treatment, followed by the two assessment posttests. All posttests were administered in paper and pencil format. Participants were instructed to proceed as quickly as possible and to refrain from editing previous items[2]. To minimize exposure to the targeted structure, production tests were administered first. Participant think-alouds throughout the experiment were recorded using Apple's Quick Time Pro software. Their mouse-click histories during the treatment were recorded by the ColdFusion application developed for this study.

Scoring

Each critical item of the recognition and production tests was worth 1 point. Participants received 1 point if they chose the correct answer and 0 points otherwise. As in Rosa and Leow (2004), agreement errors in production were not considered if intended use of the past perfect subjunctive was evident. Two raters scored all tests independently, with 98% inter-rater reliability. For the remaining 2% of mismatches, an agreed-upon solution was found.

Coding of Treatment

The mouse-clicking histories of each participant during the treatment were coded for accuracy of response, clicking order, and relative position of the clicked option in its list. As stated earlier (see *Experimental task*), similarly to Rosa and Leow (2004), two of the four options in the treatment were never the correct answers (henceforth impossible options, IMPs). Accordingly, four click types were distinguished:

- [RIGHT]: If the clicked option was correct (pluperfect subjunctive in critical items; imperfect subjunctive in distractors).
- [WRONG]: If the clicked option was incorrect but possible (imperfect subjunctive in critical items; pluperfect subjunctive in distractors).
- [IMPa]: If the clicked option pertained to the first type of impossible options (ill-formed pluperfect subjunctive or well-formed perfect conditional).
- [IMPb]: If the clicked option pertained to the second type of impossible options (well-formed simple conditional).

When participants clicked on several options before responding correctly, their successive mouse-clicks were coded from left to right. The relative position of each clicked option in its list was coded 1 to 4 (from top to bottom). For example, the sequence [(2) WRONG > (3) IMPb > (4) IMPa > (1) RIGHT] indicates that in order to fill in the blank successfully, the participant clicked on all four available options, starting with the WRONG one (which was second in the list), then clicking on IMPb (which was third), then IMPa (which was fourth), and finally the RIGHT one (which was first in the list).

Analysis

Item reliability analyses using Cronbach's alpha were performed on all assessment tests. The reliability coefficients for the recognition pretests and posttests were .82 and .88, respectively, and the coefficient for the production posttest was .95 (no analyses were necessary for the production pretests since only participants scoring 0 on all items were admitted into the study). These coefficients were satisfactory, closely approaching or passing the .9 benchmark for highly reliable research (Nunnally & Bernstein, 1994).

To probe the two research questions (i.e., the effects of type and amount of computerized input-based practice on L2 development), the raw scores from the recognition and production tests were submitted to two independent 2x2 repeated-measures ANOVAs, entering time as the within-subject factor and type or amount of input-based practice as the between-subject factor (as discussed in the *Results* section, this

statistical model was chosen based on distribution of the data). ANOVA results were interpreted in light of the observed power (*OP*) and effect size (the magnitude of the impact of the independent variable on the dependent variable). As suggested by Volker (2006), in order to provide a more fine-grained analysis both the effect size of the omnibus test effects (i.e., for the whole ANOVA—partial eta-squared, η_p^2) and the post-hoc contrast effect sizes (pairwise group comparisons— standardized mean difference effects, Cohen's *d*) were reported. An *OP* of .8 was considered acceptable. For η_p^2, .01 was considered small, .06 medium, and .14 large. For *d*, .40 was considered small, .70 medium, and 1.00 large[3]. The alpha level for all analyses of significance was set at .05.

RESULTS

Prior to running statistical analyses, the raw data on the assessment tests were checked for outliers following the outlier labeling rule in Hoaglin and Iglewicz (1987), and no outliers were identified. Visual inspection of the descriptive statistics in Table 3 and the plotted means in Figures 1 and 2 indicated that all four experimental conditions experienced learning of the targeted structure in both recognition and production tests. Combined gains were larger in recognition tests (13.96 out of 20 items) than production tests (11.44) but between-group differences were larger in the latter (4.46 items between the highest- and lowest-gaining groups versus 2 items in recognition tests). The top gainers in both tests were the 56-item groups. The ranking of groups in terms of gains was the same for both tests: 2x56 (14.69 in recognition and 14.15 in production) > 4x56 (14.54 and 11.92) > 4x28 (13.93 and 10) > 2x28 (12.69 and 9.69).

Table 3. *Descriptive Statistics: All Groups and Tests*

		Recognition				Production			
		Pretest		Posttest		Pretest		Posttest	
Group	*N*	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*
[2 x 28]	13	2.00	.81	14.69	5.76	.00	.00	9.69	7.88
[4 x 28]	13	1.92	.86	15.85	3.91	.00	.00	10.00	7.91
[2 x 56]	13	2.08	.86	16.77	3.51	.00	.00	14.15	5.11
[4 x 56]	13	1.92	.86	16.46	4.57	.00	.00	11.92	7.20
Total	52	1.98	.82	15.94	4.46	.00	.00	11.44	7.14

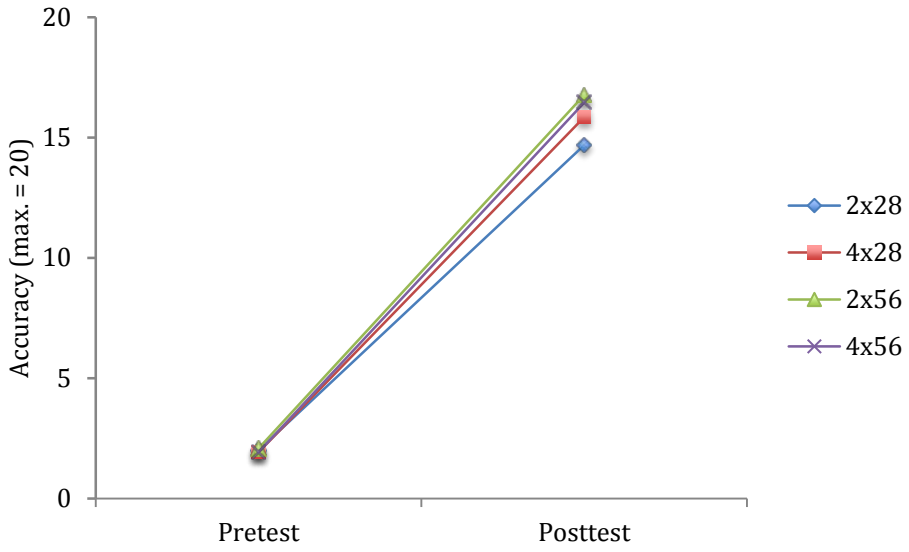

Figure 1. Written recognition accuracy by group

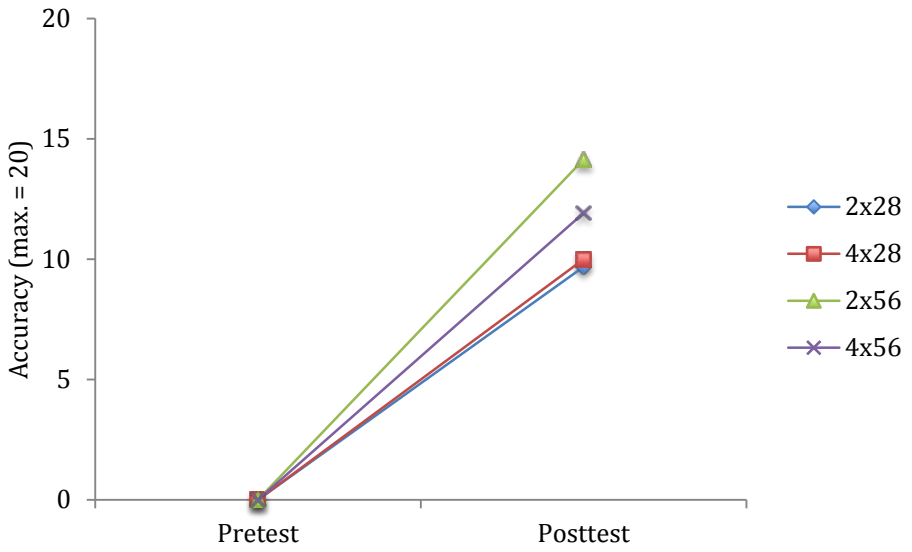

Figure 2. Written production accuracy by group

Before running parametric statistics to probe any significant effects or interactions in the results, the data were checked to ensure that assumptions of ANOVAs were met. Visual inspection of the distribution of the data in the Q-Q plots and acceptable skewness and kurtosis values (< absolute 1) indicated that the recognition and production scores in all four groups were normally distributed. Additionally, Levene's test of equality of error variances indicated that the variance between groups was not statistically different, either for recognition $F(3, 51) = .65$, $p = .587$ or production $F(3, 51) = .06$, $p = .978$[4]. To observe whether there were any differences between groups prior to the treatment, a one-way ANOVA was run on the raw scores from the recognition pretests. No analysis was necessary for the production pretests, since only participants scoring 0 on all items were admitted into the study. The results of the

ANOVA indicated that there were no statistically significant differences between groups, $F(3, 48) = .09$, $p = .96$.

RQ1. Effects of Type of Input-Based Practice

With regard to the first research question (effects of type of input-based practice), the ANOVA performed on the scores from the recognition tests yielded a significant main effect for time, $F(1, 50) = 470.07$, $p < .001$, $\eta_p^2 = .90$, no significant main effect for type of input-based practice, $F(1, 50) = .06$, $p = .807$, $\eta_p^2 = .00$, and no significant interaction Time × Type of input-based practice, $F(1, 50) = .17$, $p = .678$, $\eta_p^2 = .00$ (see Table 4). Similarly, the ANOVA performed on the scores from the production tests yielded a main effect for time, $F(1, 50) = 131.49$, $p < .001$, $\eta_p^2 = .72$, no significant main effect for type of input-based practice, $F(1, 50) = .23$, $p = .632$, $\eta_p^2 = .00$, and no significant interaction Time × Type of input-based practice, $F(1, 50) = .23$, $p = .632$, $\eta_p^2 = .00$.

Table 4. *RQ1 - Summary of ANOVA Results for Accuracy in Written Recognition and Production: Time by Type of Input-Based Practice*

Source	*df*	*SS*	*MS*	*F*	*p**	*PES*	*OP*
Written recognition							
Time	1	5068.04	5068.04	470.07	<.001	.90	1.00
Type	1	.61	.61	.06	.807	.00	.06
Time × Type	1	1.88	1.88	.17	.678	.00	.07
Written production							
Time	1	3404.09	3404.09	131.49	<.001	.72	1.00
Type	1	6.01	6.01	.23	.632	.00	.08
Time × Type	1	6.01	6.01	.23	.632	.00	.08

*Note. * alpha = .05*

The significant main effects for time and the no significant interactions Time × Type of input-based practice in both recognition and production tests indicated that the 2- and 4-option groups experienced significant and similar learning of the targeted structure in both dependent measures. However, the lack of a statistical difference must be interpreted with caution due to the very low observed power (*OP* = .07 and .08) that resulted from the sample size ($n = 13$). As Oswald and Plonsky (2010, p. 86) noted:

> Researchers have a natural tendency to interpret significant statistics no matter what the sample size is […]. This practice should generally be avoided because conceptually, small samples usually represent very little of the population of interest, and empirically, small-sample statistics (even significant ones) are highly unstable.

Along these lines, Neill (2008) noted that significance testing with small sample sizes can be misleading because it is subject to Type II errors (i.e., concluding that there is no effect of a treatment when there is in fact one), and suggested using means and standardized mean difference effect sizes as more informative measures. Following this suggestion, the standardized mean difference effect sizes between the 2- and 4-option groups were computed and compared. For recognition, d([2x28] × [4x28]) = .24 and d([2x56] × [4x56]) = .08. For production, d([2x28] × [4x28]) = .04 and d([2x56] × [4x56]) = .36. Based on Oswald and Plonsky's (2010, p. 99) standards for effect sizes, only the latter comparison ([2x56] × [4x56]) approached a small effect (i.e., $d = 36$).

To further elucidate these results, the participant clicking histories in the 4-option groups (4x28 and 4x56) were analyzed. Figure 3 shows a visual representation of the total number of clicks on IMPs for the shared first 28 items. The white and black bars represent critical items (eliciting the targeted structure) and distractors (eliciting the non-targeted structure), respectively. As can be observed, all the peaks in the graph (i.e., the most frequently clicked options) correspond to distractors. This indicates that participants completed the experimental task in the following way: if the item was critical, participants narrowed down their choices to the two relevant options and rarely clicked on IMPs; however, if the item was a distractor, they clicked on the IMP options more often. Observation of the clicking order revealed that participants usually clicked on the targeted structure first, likely because critical items were more frequent (18 out of 28), but when they received negative feedback in a distractor they tried more options, including the IMPs.

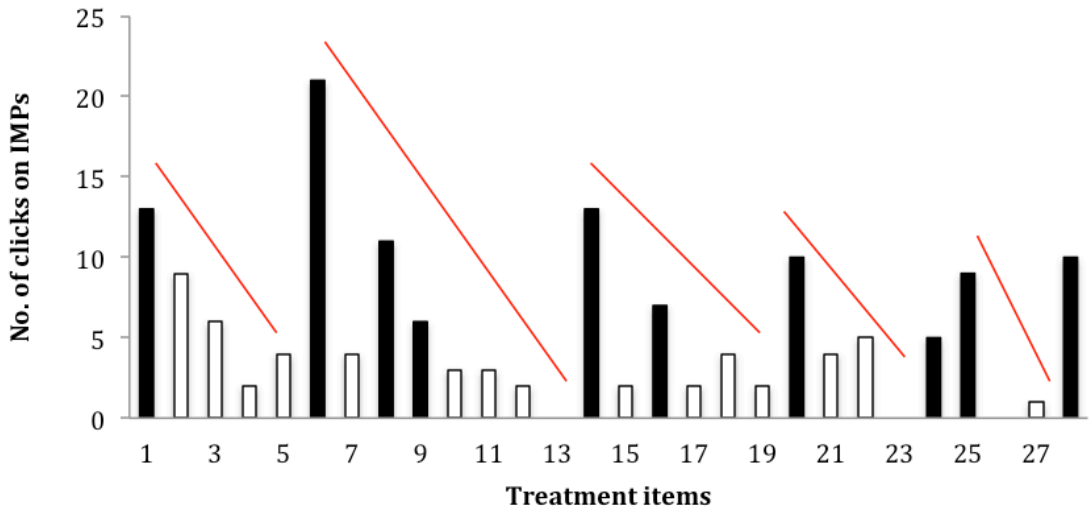

Figure 3. Number of clicks on IMPs by treatment item, including critical items (white bars), distractors (black bars), and clicking trends between critical items and distractors (red lines) (4x28 and 4x56 groups combined).

Let us now look at the individual participants who clicked on IMPs, rather than the total number of clicks on IMPs. Figure 4 provides a visual representation of the number of participants in the 4x56 group (n = 13) who clicked on IMPs throughout the treatment. As this graph illustrates, the number of participants clicking on IMPs ranged from zero (critical items 4, 13, 23, 26, 29, 34, 35, 36, 44, 46, 53) to six (distractor item 6). On average, only 1 out of 13 (7.7%) participants clicked on an IMP if the item was critical, while this number tripled (23%) if the item was a distractor. Dividing the treatment into its two halves does not change this pattern, but it reveals that participants gradually clicked on fewer IMPs. For critical items, the average number of participants who clicked on IMPs went from 1.3 (10%) in the first half to 0.8 (6%) in the second. For distractors, the number shifted from 3.5 (27%) to 2.6 (20%). Moreover, during the second half of the treatment (items 29–56) 92% of the participants clicked on IMPs on three occasions or fewer, which indicates that most of the participants identified the relevant options with great rates of success.

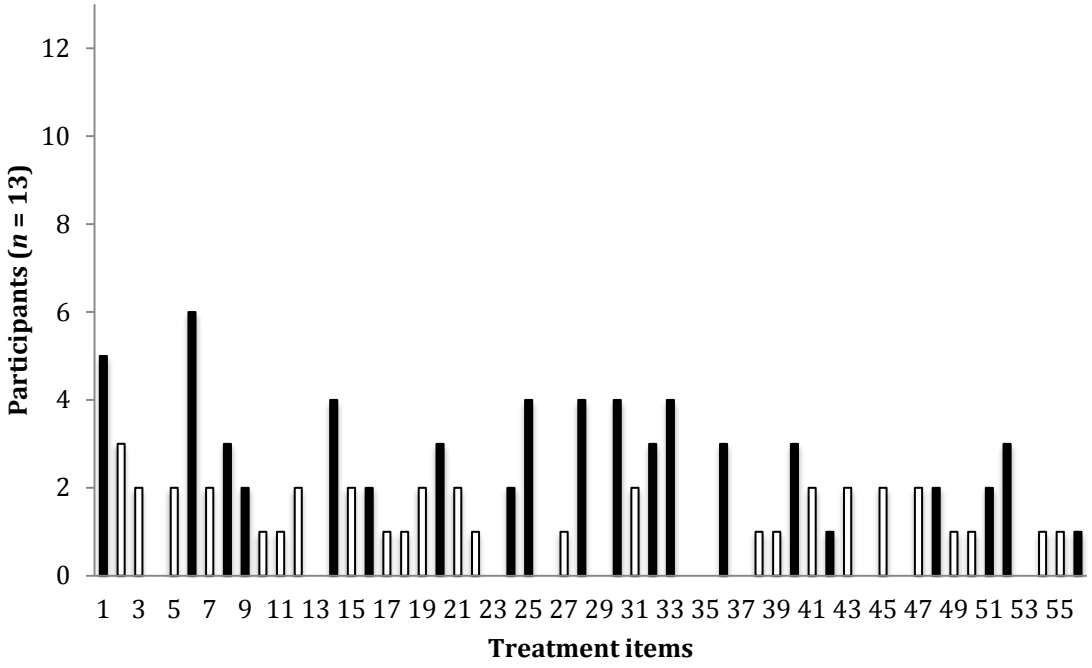

Figure 4. Number of participants clicking on IMPs by treatment item, including critical items (white bars) and distractors (black bars) (4x56 group)

To sum up, based on the results from the ANOVAs and the standardized mean difference effect sizes it seems that the different numbers of options in the computerized input-based practice in Rosa and Leow (2004) and Sanz and Morgan-Short (2004) did not have a differential effect on the development of the L2 targeted form. Qualitative observation of participant mouse-click histories revealed that participants clicked on the additional options in Rosa and Leow very rarely, mostly upon encountering a distractor, and progressively fewer times as the treatment evolved.

RQ2. Effects of Amount of Input-Based Practice

With regard to the second research question (effects of amount of input-based practice), the ANOVA performed on the scores from the recognition tests yielded a significant main effect for time, $F(1, 50) = 478.26$, $p < .001$, $\eta_p^2 = .90$, no significant main effect for amount of input-based practice, $F(1,50) = 1.25$, $p = .268$, $\eta_p^2 = .02$, and no significant interaction Time × Amount of input-based practice, $F(1, 50) = 1.05$, $p = .311$, $\eta_p^2 = .02$ (see Table 5). Again, similar results were obtained for production, with a significant main effect for time, $F(1, 50) = 137.91$, $p < .001$, $\eta_p^2 = .73$, no significant main effect for amount of input-based practice, $F(1, 50) = 2.68$, $p = .108$, $\eta_p^2 = .05$, and no significant interaction Time × Amount of input-based practice, $F(1, 50) = 2.68$, $p = .108$, $\eta_p^2 = .05$.

Table 5. *RQ2 - Summary of ANOVA Results for Accuracy in Written Recognition and Production: Time by Amount of Practice*

Source	*df*	*SS*	*MS*	*F*	*p**	*PES*	*OP*
Written recognition							
Time	1	5068.04	5068.04	478.26	<.001	.90	1.00
Amount	1	12.46	12.46	1.25	.268	.02	.19
Time × Amount	1	11.11	11.11	1.05	.311	.02	.17
Written production							
Time	1	3404.09	3404.09	137.91	<.001	.73	1.00
Amount	1	66.24	66.24	2.68	.108	.05	.36
Time × Amount	1	66.24	66.24	2.68	.108	.05	.36

*Note. * alpha = .05*

The significant main effects for time and the no significant interactions Time × Amount of input-based practice in both recognition and production tests indicated that the 28- and 56-item groups experienced significant and similar learning of the targeted structure in both measures. These results must again be interpreted with caution due to the low observed power (*OP* = .17 and .36). In fact, the small (η_p^2 = .02) to almost medium (η_p^2 = .05) effect sizes of the no significant interactions in recognition and production tests suggest that the larger gain differences of the 56-item groups (1.34 and 3.19 more items in recognition and production) are still worth noting. To probe deeper into these results, the standardized mean difference effect sizes between the 28- and 56-item groups were computed and compared. For recognition, *d*([2x28] × [2x56]) = .44 and *d*([4x28] × [4x56]) = .14. For production, *d*([2x28] × [2x56]) = .67 and *d*([4x28] × [4x56]) = .25. According to Oswald and Plonsky's (2010, p. 99) standards, the standardized mean difference effect sizes between the 2x28 and 2x56 groups reached a small effect in recognition and a medium effect in production. Consequently, despite the finding of a no significant interaction Time × Amount of input-based practice, the larger mean gains of the 56-item groups and the observed small to medium effect sizes for the 2-option groups suggested that the higher amount of practice items completed by the *right/wrong* feedback group in Sanz and Morgan-Short (2004) vis-à-vis Rosa and Leow (2004) might have added some edge in the development of the targeted form, particularly in production.

DISCUSSION

The first research question asked whether type of computerized input-based task, with either two or four options, as in Sanz and Morgan-Short (2004) and Rosa and Leow (2004), respectively, had differential effects on L2 development. Since the present study was the first to investigate this issue, the null hypothesis was formulated. Quantitative results failed to reject the null hypothesis, with the 2- and 4-option groups experiencing comparable learning achievement in recognition and production of the targeted form. Qualitative observation of participant mouse-click histories revealed that participants clicked on the additional options in the 4-option groups very rarely, mostly upon encountering a distractor, and progressively fewer times as the treatment evolved.

To investigate why this happened, the participant think-alouds during the treatment phase were monitored. This revealed that participants were able to identify the targeted structure since very early in the treatment. For example, on item 6 one participant verbalized:

> *Mi hermana Isabel sabe hablar italiano… Si* Isabel… uh… I think it has to be past subjunctive… there's only one past subjunctive that I see…

As the treatment progressed, participants made even more explicit statements about the only two relevant options. For example, on item 42, another participant stated:

> *Quiero viajar*, I wanna like travel to Korea, but my [breath] Korean friend doesn't invite me. [Tsk] [breath] I would travel to Korea… South Korea that is [breath] if *mi amiga*, if my friend… ah… had invited me. *Hubiera invitado*. Ooops. *Invitara* [breath]. There are only two that are past subjunctive, max on these.

Therefore, the present study does not provide empirical support to Sanz and Morgan-Short's (2004, p, 71) claim that the presence of additional options in Rosa and Leow's (2004) experimental task may have impeded participants from noticing the target form. Evidence from participant mouse-click histories and think-aloud protocols showed that participants completing the 4-option task were able to identify the two relevant options (targeted and non-targeted forms) rather quickly through an elimination process. This is consistent with Leow (2000) and Hama and Leow (2010), in which think-aloud protocols gathered on the post-exposure assessment tasks revealed that some learners engaged in the strategy of eliminating several distracting options to arrive at the final one. Also, these findings underscore the role of learners as active agents in the learning process and the disconnect between what Ellis (2003) and Seedhouse (2005) call *task-as-a-workplan* (i.e., what teachers/researchers think the task will do) and *task-as-a-process* (i.e., what learners actually do) or, in terms of sociocultural and activity theorists (e.g., Coughlan & Duff, 1994), the disconnect between *tasks* and *activities* (see, e.g., Dooly, 2011 for recent empirical evidence).

The second research question asked whether amount of input-based practice (28 items, as in Rosa & Leow, 2004 versus 56 items, as in Sanz & Morgan-Short, 2004) had differential effects on L2 development. Based on previous empirical literature, it was hypothesized that learners completing 56 items of a computerized task with *right/wrong* feedback would outperform those completing 28 items. While the larger gains of the 56-item groups in recognition (1.34 more items) and production (3.19 more items) did not reach statistical significance, the small to medium effect sizes observed suggested that these differences may reach significance with a larger population, particularly in production.

On the one hand, these results provide partial support to the growing body of literature showing that task-repetition in combination with corrective feedback has beneficial effects on accuracy (Adams, 2003; Sheppard, 2006; Leow, 1998). The fact that the extra practice helped more in production than recognition is probably because input-processing (comprehension, recognition, or interpretation) is less challenging than output production (see, e.g., Flynn, 1986), and thus 28 items sufficed to attain an already high level of recognition accuracy (76% versus 49% production accuracy).

On the other hand, it is surprising that the extra practice did not have larger effects, particularly because there was room for improvement (3.4 and 6.96 items in recognition and production posttests, respectively). Based on informal post-debriefing interviews with the participants, it was posited that the extra practice might not have produced the expected effects because it induced fatigue and/or boredom. Indeed, the think-alouds in the 56-item groups provided many instances of participants who complained about the large number of items, both during the treatment and posttests. Observe, for example, the following verbalizations of participant Pupi (not her real name):

> [Treatment item #22:] *Las vacaciones se han terminado. Muchos estudiantes estarían todavía en casa de su familia si las vacaciones se… de su familia… terminaran?* Huh? *Hubieran terminado*? [stretching] I'm bored!

> [Treatment item #46:] *habría limpiado la casa si*… If he had time? No? Uh…. How many more left? [sigh] Uh…
>
> [Treatment item #51, addressing her neighbor:] How many, how many are on here? I'm on like 51! Is that right? [Neighbor:] Mine just stopped and said thank you for completing the survey. [Pupi:] OH MY GOD!!! [laughter] This is not ending!!! [laughter] [Neighbor:] Yeah, mine didn't have that many [Pupi:] Erm… [addressing one researcher] Hi, er… how many questions are there? [Researcher:] Oh, you're almost done, you're almost done [laughter]. [Pupi:] Ok [Researcher:] Yeah, yeah [Pupi:] I was like wondering [Researcher laughs].
>
> [Treatment item #56:] Oh God!
>
> [Production posttest item #12, after turning the page and discovering more questions:] Oh, dammit! [sigh] [puff]

In their study, Sanz and Morgan-Short (2004) claimed: "Perhaps if the amount of practice in [Rosa and Leow (2004)] had been greater, the [*right/wrong* feedback] group would not have performed significantly [worse than] the other groups." While the present study suggests that this is possible, it also underscores that 56 items in one batch may induce fatigue and/or boredom, which may in turn decrease the effects of the extra practice. Maybe if the extra practice had been administered in different sessions, with time lapses in between, as in Leow (1998), higher accuracy gains would have been observed.

Taken together, the results from the present study suggest that of the two variables under investigation, type and amount of input-based practice, only the latter may possibly contribute to explaining why the *right/wrong* feedback group in Sanz and Morgan-Short attained learning gains comparable to their feedback group with grammatical explanations. However, the small to medium effect sizes observed here (d = .44 and .67) do not make up for the large effect sizes observed in Rosa and Leow (2004). The standardized mean difference effect sizes between their *right/wrong* feedback group and the group that received feedback with grammatical explanations reached d values of 1.31 and 1.69 in recognition of old and new items and 2.68 and 1.36 in production of old and new items, respectively. Hence, there must be additional variables, other than amount of practice, that explain the conflicting findings between Rosa and Leow and Sanz and Morgan-Short (2004), as explained in the next section.

LIMITATIONS AND FURTHER RESEARCH

As with all studies, there are certain methodological limitations that should be noted. First, the number of participants per group was relatively low (n = 13). Perhaps if the sample size had been larger the mean differences between the 28- and 56-item groups would have reached significance. Second, assessment measures included immediate written recognition and production only, did not include a "Don't know" option, and were administered in printed form. Future studies may want to include delayed posttests using a wider variety of dependent measures, assess whether or not learners were guessing, and preclude backtracking more effectively. And third, while the use of mouse-click tracking and think-alouds proved very valuable to uncover the learning behavior of the participants, it may be more effective to record both of them under one single (synchronized) track via multimedia screen-capture software.

The present study also opened some avenues for further research. First, future studies may want to investigate the effects of additional practice in several batches to diminish the potentially detrimental effects of fatigue or boredom. Second, it would be very interesting to use effective ways of operationalizing and measuring fatigue and boredom and analyze their potential effects on L2 development through correlational research. Think-alouds offer valuable information but are not ideal, because not all participants verbalize their feelings. Perhaps these data could be triangulated with surveys

and more sophisticated measures of fatigue, like electroencephalographic measures and response tests in visual display terminal tasks (Cheng & Hsu, 2011). Third, differences in amount of practice, as suggested earlier, may not be solely responsible for the conflicting findings between the two studies compared here (and the remaining studies on the effects of type of computerized feedback). The picture interpretation task in Sanz and Morgan-Short (2004) might have promoted stronger form-meaning connections than the merely textual multiple-choice task in Rosa and Leow (2004). Future studies should investigate the effects of different types of input-based tasks beyond their number of options. Finally, type of targeted structure may also exert a moderating role. The studies that found superior performance for feedback with grammatical explanations targeted Japanese particles and passivization (Nagata, 1993), Japanese reflexive constructions (Sachs, 2011), Latin assignment of semantic functions (Lado et al., 2014; Lin, 2009), Spanish dative experiencer constructions with *gustar* (Bowles, 2008), Spanish past counterfactual conditional sentences (Rosa & Leow, 2004), and Spanish present subjunctive and preposition pied-piping in adjectival relative clauses (Cerezo, 2010). In contrast, the studies that found comparable performance for *right/wrong* feedback targeted Spanish noun-adjective gender and number agreement (Camblor, 2006), Spanish direct object pronouns (Moreno, 2007; Sanz & Morgan-Short, 2004), and Spanish dative experiencer constructions with *gustar* (Hsieh, 2008). Hulstijn and de Graaff (1994) argued that metalinguistic explanations give an edge when the targeted structures are more complex. Perhaps the structures in Rosa and Leow and the studies that align with it were more complex than those in the group of studies like Sanz and Morgan-Short (though see Bowles, 2008 vs. Hsieh, 2008 for *gustar*). Clearly, future studies must investigate the effects of different types of computerized feedback on different targeted structures (see, e.g., Cerezo, 2010).

CONCLUSIONS

The theoretical and methodological contributions of this study have implications for research on, and the development of, the ever more present hybrid and online L2 learning programs. From a theoretical standpoint, this study suggests that the conflicting findings in the literature on type of computerized feedback (the effects of *right/wrong* feedback versus feedback with grammatical explanations) might be due to differences in experimental treatments. Addressing claims by Sanz and Morgan-Short (2004) and partially replicating Rosa and Leow (2004), this study concluded that the superior performance of the feedback with grammatical explanations in Rosa and Leow might have been due to the inferior amount of computerized practice provided, which may have put the *right/wrong* feedback group at a disadvantage. Possibly, however, this was not the only reason. Differences in the type of practice and the targeted structure may also have contributed to these conflicting findings.

From a methodological perspective, this study underscored the importance of triangulating data. Measures of online behavior (mouse-click tracking and think-alouds) were helpful to uncover learner strategies evidencing the mismatch between the notions of *task-as-a-workplan* and *task-as-a-process*, as well as the increased boredom and fatigue induced by the extra amount of practice. Similarly, the use of descriptive statistics and standardized mean difference effect sizes proved very useful to reinterpret the results of significance tests, which in the case of small sample sizes may propel Type II errors.

APPENDIX. Examples of a critical item and distractor in the experimental task

Critical item (English translation)

Context: "The Wright brothers built the first airplane."

Sentence: "We would not be able to travel to Europe so fast these days if the Wright brothers _____ the first airplane."

Options: hadn't built

wouldn't build

didn't build

wouldn't have built

Distractor (English translation)

Context: "My sister Isabel can speak Italian."

Sentence: "If she _____ her next vacation in Italy, she would be able to communicate with the natives."

Options: would have spent

had spent

spent

would spend

NOTES

1. An anonymous reviewer pointed out that the absence of a "Not sure/Don't know" option in the recognition task "could have potentially skewed results since students were forced to pick an option even if that meant they needed to guess." While this is certainly possible, including such option would have altered the partial replication nature of this study, as neither Sanz and Morgan Short (2004) nor Rosa and Leow (2004) included it. Future studies, however, need to address the role of "Not sure/Don't know" options in input-based tasks.

2. An anonymous reviewer suggested that the paper-and-paper pencil nature of the assessment tests does not guarantee that participants did not backtrack, contrary to the instructions they received. This is possible but unlikely because the laboratory was constantly proctored by the researcher and two assistants. Also, the think-alouds did not reveal any cases of backtracking.

3. The coefficients for *OP* and η_p^2 were based on Cohen (1988). The coefficients for Cohen's *d* were based on Oswald and Plonsky's (2010, p. 99) revision of Cohen's (1988) benchmarks due to their more conservative nature (they require an additional .20 for every category). It should be noted, however, that these revised coefficients constitute "a preliminary and general set of SLA standards for effect sizes" (Oswald & Plonsky, 2010, p. 99).

4. Tests of normality of distribution cannot be interpreted with reliable confidence with small sample sizes. For that reason, statistical analyses were also performed with a nonparametric statistical model (Kruskal-Wallis on learning gains). Results did not differ from the ANOVA analyses reported here.

ACKNOWLEDGEMENTS

I would like to thank Rebecca Sachs, Alison Mackey, and Donna Lardiere for their comments on earlier versions of this paper. I am also very grateful to the three anonymous reviewers for their helpful feedback, Roberto Gómez Fernández and Boram Suh for their help with data gathering, and the students who participated in this study. All errors are of course my own.

ABOUT THE AUTHOR

Luis Cerezo is Assistant Professor of Spanish Applied Linguistics and Director of the Spanish Language Program at American University (Washington, DC). Author of various publications on hybrid and online language learning, he developed *Talking to Avatars*, an e-tutor in which learners interact with pre-filmed actors to learn Spanish in real-life situations.

E-mail: cerezoce@american.edu

REFERENCES

Adams, R. (2003). L2 output, reformulation and noticing: Implications for IL development. *Language Teaching Research*, *7*(3), 347–376.

Ahmadian, M. J. (2011). The effect of 'massed' task repetitions on complexity, accuracy and fluency: Does it transfer to a new task? *Language Learning Journal*, *39*(3), 269–280.

Ahmadian, M. J., & Tavakoli, M. (2011). The effects of simultaneous use of careful online planning and task repetition on accuracy, complexity, and fluency in EFL learners' oral production. *Language Teaching Research*, *15*(1), 35–59.

Bowles, M. A. (2008). Task type and reactivity of verbal reports in SLA: A first look at a L2 task other than reading. *Studies in Second Language Acquisition*, *30*(3), 359–387.

Bygate, M. (1996). Effect of task repetition: Appraising the development of second language learners. In J. Willis & D. Willis (Eds.), *Challenge and change in language teaching* (pp. 136–146). Oxford: Heinemann.

Bygate, M. (2001). Effects of task repetition on the structure and control of oral language. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Researching pedagogic tasks, second language learning, teaching and testing* (pp. 23–48). Harlow, UK: Longman.

Camblor, M. (2006). *Type of written feedback, awareness, and L2 development: A computer-based study.* (Unpublished doctoral dissertation). Georgetown University, Washington, DC

Cerezo, L. (2010). *Talking to avatars: The computer as a tutor and the incidence of learner's agency, feedback, and grammatical form in SLA.* (Unpublished doctoral dissertation). Georgetown University, Washington, DC.

Cerezo, L. (2012). Beyond hybrid learning: A synthesis of research on e-tutors under the lens of SLA theory. In F. Rubio & J. J. Thoms (Eds.), *Hybrid language teaching and learning: Exploring theoretical, pedagogical and curricular issues* (pp. 50–66). Boston: Heinle/Cengage Learning.

Cheng, S. Y., & Hsu, H. T. (2011). Mental fatigue measurement using EEG. In G. Nota (Ed.), Risk management trends. Retrieved from http://www.intechopen.com/books/risk-management-trends/mental-fatigue-measurement-using-eeg

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Coughlan, P., & Duff, P. (1994). Same task, different activities: analysis of SLA from an activity theory perspective. In J. Lantolf & G. Appel (Eds.), *Vygotskian approaches to second language research* (pp. 173–194). Norwood, NJ: Ablex.

Craik, F. I. M., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General*, *104*(3), 268–294.

Dooly, M. (2011). Divergent perceptions of telecollaborative language learning tasks: Tasks-as-workplan vs. task-as-process. *Language Learning & Technology*, *15*(2), 69–91. Retrieved from http://llt.msu.edu/issues/june2011/dooly.pdf

Ellis, R. (2003). *Task-based language learning and teaching*. Oxford: Oxford University Press.

Ellis, R. (2009). The differential effects of three types of task planning on the fluency, complexity, and accuracy in L2 oral production. *Applied Linguistics*, *19*(4), 474–509.

Flynn, S. (1986). Production vs. comprehension: Differences in underlying competences. *Studies in Second Language Acquisition*, *8*(2), 135–164.

Gass, S. M., Mackey, A., Fernández, M., & Álvarez-Torres, M. J. (1999). The effects of task repetition on linguistic output. *Language Learning*, *49*(4), 549–580.

Hama, M., & Leow, R. P. (2010). Learning without awareness revisited: Extending Williams (2005). *Studies in Second Language Acquisition*, *32*(3), 465–491.

Hawkes, M. L. (2012). Using task repetition to direct learner attention and focus on form. *ELT Journal*, *66*(3), 327–336.

Hoaglin, D. C., & Iglewicz, B. (1987). Fine tuning some resistant rules for outlier labeling. *Journal of the American Statistical Association*, *82*(400), 1147–1149.

Hsieh, H. C. (2008). The effects of type of exposure and type of post-exposure task on L2 development. *Journal of Foreign Language Instruction*, *2*(1), 117–138.

Hulstijn, J. H., & de Graaff, R. (1994). Under which conditions does explicit knowledge of a second language facilitate the acquisition of implicit knowledge? A research proposal. *AILA Review*, *11*, 97–112.

Lado, B., Bowden, H., Stafford, C., & Sanz, C. (2014). A fine-grained analysis of the effects of negative evidence with and without metalinguistic information in language development. *Language Teaching Research*, *18*(3), 320–344.

Larsen-Freeman, D. (2009). Adjusting expectations: The study of complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics*, *30*(4), 579–589.

Leow, R. P. (1998). The effects of amount and type of exposure on adult learners' L2 development in SLA. *The Modern Language Journal*, *82*(1), 49–68.

Leow, R. P. (2000). A study of the role of awareness in foreign language behavior: Aware vs. unaware learners. *Studies in Second Language Acquisition*, *22*(4), 557–584.

Levelt, W. (1989). *Speaking: From intention to articulation*. Cambridge, MA: MIT Press.

Levy, M. (2009). Technologies in use for second language learning. *The Modern Language Journal*, *93*(Focus Issue), 769–782.

Li, S. (2010). The effectiveness of corrective feedback in SLA: A meta-analysis. *Language Learning*, *60*(2), 309–365.

Lin, H. J. (2009). *Bilingualism, feedback, cognitive capacity, and learning strategies in L3 development.* (Unpublished doctoral dissertation). Georgetown University, Washington, DC.

Lynch, T., & Maclean, J. (2000). Exploring the benefits of task repetition and recycling for classroom language learning. *Language Teaching Research*, *4*(3), 221–250.

Lynch, T., & Maclean, J. (2001). Effects of immediate task repetition on learners' performance. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Researching pedagogic tasks, second language learning, teaching and testing* (pp. 99–118). Harlow, UK: Longman.

Lyster, R., & Saito, K. (2010). Oral feedback in classroom SLA: A meta-analysis. *Studies in Second Language Acquisition*, *32*(2), 265–302.

Mackey, A., & Goo, J. (2007). Interaction research in SLA: A meta-analysis and research synthesis. In A. Mackey (Ed.), *Conversational interaction in second language acquisition: A collection of empirical studies* (pp. 407–452). Oxford, UK: Oxford University Press.

Medina, A. D. (2008). *Concurrent verbalization, task complexity, and working memory: Effects on L2 learning in a computerized task.* (Unpublished doctoral dissertation). Georgetown University, Washington, DC.

Moreno, N. (2007). *The effects of type of task and type of feedback on L2 development in CALL.* (Unpublished doctoral dissertation). Georgetown University, Washington, DC.

Morgan-Short, K., & Wood Bowden, H. (2006). Processing instruction and meaningful output-based instruction: Effects on second language development. *Studies in Second Language Acquisition*, *28*(1), 31–65.

Nagata, N. (1993). Intelligent computer feedback for second language instruction. *Modern Language Journal*, *77*(3), 330–339.

Nagata, N. (1998a). Input vs. output practice in educational software for second language acquisition. *Language Learning & Technology*, *1*(2), 23–40. Retrieved from http://www.llt.msu.edu/vol1num2/pdf/article1.pdf.

Nagata, N. (1998b). The relative effectiveness of production and comprehension practice in second language acquisition. *Computer Assisted Language Learning*, *11*(2), 153–177.

Nagata, N., & Swisher, M. V. (1995). A study of consciousness-raising by computer: The effect of metalinguistic feedback on SLA. *Foreign Language Annals*, *28*(3), 336–347.

Neill, J. (2008). Why use effect sizes instead of significance testing in program evaluation. Retrieved from http://wilderdom.com/research/effectsizes.html

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric Theory* (3rd ed.). New York, NY: McGraw-Hill.

Oswald, F. L., & Plonsky, L. (2010). Meta-analysis in second language research: Choices and challenges. *Annual Review of Applied Linguistics*, *30*, 85–110.

Rosa, E. M., & Leow, R. P. (2004). Computerized task-based exposure, explicitness, type of feedback, and Spanish L2 development. *The Modern Language Journal*, *88*(2), 192–216.

Russell, J., & Spada, N. (2006). The effectiveness of corrective feedback for the acquisition of L2 grammar. In J. M. Norris (Ed.), *Synthesizing research on language learning and teaching* (pp. 133–162). Philadelphia, PA: John Benjamins.

Sachs, R. (2011). *Individual differences and the effectiveness of visual feedback on reflexive binding in L2 Japanese.* (Unpublished doctoral dissertation). Georgetown University, Washington, DC.

Sanz, C., Morales-Front, A., Zalbidea, J., & Zárate-Sández, G. (2015). Always in motion is the future: Doctoral students' use of technology for SLA research. In R. P. Leow, L. Cerezo, & M. Baralt (Eds.), *A Psycholinguistic Approach to Technology and Language Learning* (pp. 49–68). New York: De Gruyter Mouton.

Sanz, C., & Morgan-Short, K. (2004). Positive evidence versus explicit rule presentation and explicit negative feedback: A computer-assisted study. *Language Learning*, *54*(1), 35–78.

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

Seedhouse, P. (2005). "Task" as a research construct. *Language Learning*, *55*(3), 533–570.

Sheppard, C. (2006). *The effects of instruction directed at the gaps second language learners noticed in their oral production.* (Unpublished doctoral dissertation). University of Auckland, Auckland, New Zealand.

Shintani, N., Li, S., & Ellis, R. (2013). Comprehension-based versus production-based instruction: A meta-analysis of comparative studies. *Language Learning*, *63*(2), 296–329.

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

Torres, J. R. (2013). *Heritage and second language learners of Spanish: The roles of task complexity and inhibitory control.* (Unpublished doctoral dissertation). Georgetown University, Washington, DC.

Volker, M. A. (2006). Reporting effect size estimates in school psychology research. *Psychology in the Schools, 43*(6), 653–672.

Wang, Z. (2009). *Modelling speech production and performance: Evidence from five types of planning and two task structures.* (Unpublished doctoral dissertation). Chinese University of Hong Kong, Hong Kong, China.