REPLICATION STUDY

Clarifying the role of inhibitory control in L2 phonological processing: A preregistered, close replication of Darcy et al. (2016)

Amanda Huensch

University of Pittsburgh
Email: amanda.huensch@pitt.edu

Abstract

Darcy et al. (2016) examined the relationship between language abilities and general cognition, or specifically, how inhibitory control might relate to L2 speech perception and production. Given that their findings unexpectedly indicated a stronger relationship between inhibitory control and perception in comparison to inhibitory control and production, and because inhibitory control was measured using a single, retrieval-induced inhibition task, the current study is a close replication with the inclusion of two additional tasks of intentional inhibition: the Stroop task and the Simon task. A comparison of the descriptive statistics for the tests of phonological processing and retrieval-induced inhibition between the initial study and current replication indicated negligible differences between the two participant samples. However, results of the partial correlation analyses in the current replication did not indicate clear relationships between phonological variables and inhibitory control. Possible explanations for the different patterning of results and implications for future replications are discussed.

Introduction

A better understanding of domain-general cognitive mechanisms (e.g., attention, working memory, inhibitory control) and their connections to language learning and use have recently garnered increased attention in the field of second language acquisition (SLA) (Luque & Morgan-Short, 2021; McManus, 2021). This is, in part, because of the central role they are hypothesized to play within usage-based approaches to SLA: Learning a language is accomplished using the same mechanisms used for learning other skills. Relatedly, scholars have sought to explain variability in language learning outcomes as a result of individual differences in these mechanisms (Botezatu et al., 2022; Saito et al., 2022). Inhibitory control, or inhibition, has surfaced as one potentially important executive function for language use, proposed as a cognitive mechanism to explain language selection in bilinguals who must continuously inhibit and/or activate relevant lexical items their competing languages (Green, 1998; Green & Abutalebi,

2013). Inhibitory control, in fact, encompasses multiple cognitive processes related to the suppression of irrelevant/distracting information and/or prepotent (i.e., dominant/ automatic) responses (Friedman & Miyake, 2004) and is therefore potentially relevant to language learning and use in multiple ways. Although a variety of taxonomies exist for conceptualizing inhibitory control (see Friedman & Miyake, 2004, for a summary), it is generally understood to consist of processes that vary according to whether they are (i) intentional/unintentional, (ii) behavioral/cognitive (i.e., a motor response or a mental process), and (iii) resisting interference from something previously learned (i.e., proactive interference) or from an immediately distracting stimulus.

Retrieval-induced inhibition has been argued to reflect one type of inhibitory control used during language processing: Through the process of lexical retrieval, non-retrieved lexical items are inhibited (Dijkstra & van Heuven, 2002; Lev-Ari & Peperkamp, 2013). For instance, if one is asked to memorize a list of vocabulary belonging to a particular lexical category (e.g., animals) and then practices a subset of words from the list, the non-practiced words will be inhibited (i.e., have their activation suppressed) such that if one is asked to recall a non-practiced word on a subsequent task, retrieval time will be slowed. Furthermore, the stronger one's individual inhibitory control, the more delayed the retrieval time would be. In this way, retrieval-induced inhibition represents unintentional, cognitive, resistance to proactive interference (Friedman & Miyake, 2004), as it operates before awareness—as opposed to active, conscious suppression—and resists the influence of prior-learned information.

Although inhibitory control has been investigated relatively extensively in the field of bilingual language processing, particularly in relation to sentence comprehension and lexical access (Costa & Santesteban, 2004; Filippi et al., 2012), Darcy, Mora, and Daidone (2016) was a pioneering study that extended these ideas to L2 phonological processing. They provided preliminary evidence that competition-based accounts of language use might explain individual differences in L2 pronunciation. Motivated in part by the work of Lev-Ari and Peperkamp (2013) demonstrating a relationship between inhibitory control and the amount of L1 attrition present in the perception and production of voiceless stops by English L1 learners of L2 French, Darcy et al. (henceforth DM&D) hypothesized that if inhibitory control could be shown to be related to L1 phonology, then it might also play a crucial role in L2 pronunciation skills: "being able to suppress the L1 more robustly could help L2 users reduce interference from their L1 segment categories during L2 use" (p. 745). As a result, variability in outcomes in L2 pronunciation could be the result of individual differences in inhibitory control.

To explore the relationship between inhibitory control and L2 phonological processing, DM&D compared the performance on a retrieval-induced inhibition task to L2 perception and production performance focusing on both consonant and vowel targets. Participants included Spanish L1 learners of English and English L1 learners of Spanish. L2 perception was measured using a speeded ABX categorization task; L2 production was measured using a delayed sentence repetition task; and L2 proficiency was measured as a function of vocabulary size using the X_Lex task (Meara, 2005). The L2 perception speeded ABX categorization task was designed such that the same stimuli could be used for each L2 group, thus allowing for generalizability while simultaneously including an internal control mechanism for the materials. DM&D predicted that greater inhibitory control would be associated with higher accuracy in both perception (more accurate identification of target sounds) and production (more accurate production of target sounds) once proficiency was controlled. Instead, their findings appeared to indicate a stronger relationship between inhibition and perception—in

comparison to production—which they suggested could be because perception relies on category formation while production additionally requires the inhibition of a dominant motor response (i.e., the articulation of an L1 segment). In other words, production skills may (additionally) rely on another category of inhibitory control—prepotent response inhibition (Friedman & Miyake, 2004), or the intentional resistance of a dominant/automatic response. Therefore, the current study is a close replication of DM&D, with the variable modification of additional tests of inhibitory control measuring prepotent response inhibition.

One of the most common measures of prepotent response inhibition is the Stroop task (Stroop, 1935), which has also been demonstrated to be related to perceptual adaptation (Kim et al., 2020) and L1 speech perception (Lev-Ari & Peperkamp, 2014). In the classic version of the Stroop task, written words of color terms (e.g., *green*, *blue*) are presented in different colors of ink which either match or not the semantics of the word (e.g., seeing the word *red* in red ink and responding "red" vs. seeing the word *red* in blue ink and responding "blue"), and participants orally respond by indicating the color of the written word. More recent iterations (Bearden et al., 2021; Gass & Lee, 2011) involve manual versions of the task in which participants are trained to associate the ink colors with different keyboard button presses. Although the Stroop task is a classic test of prepotent response inhibition and thus an ideal candidate for use in the current replication, one final consideration relates to domain general rather than language-oriented response inhibition (Lev-Ari & Peperkamp, 2014; Linck et al., 2012). The Simon task (Simon & Rudell, 1967), also a task that requires the inhibition of a prepotent response, has been argued to be non-linguistic/domain general (Linck et al., 2012) because it involves identifying the spatial location of an object (e.g., whether a box is presented on the left of right side of a computer screen) rather than ignoring linguistic interference from written words.

A final point of consideration regarding the findings of DM&D is that the results did not appear to pattern in the same way for both consonant and vowel targets. For vowels, higher inhibitory control was related to more accurate perception, but no relationship emerged between inhibitory control and production. For consonants, higher inhibitory control was related to more accurate production, but no results were able to be reported for perception because the L2 English group performed at ceiling on the perception task. Therefore, to provide a complete picture of the relationship between inhibitory control and perception and production of both consonant and vowel targets, the current replication only recruited L2 Spanish participants, as this group showed sufficient variation on the consonant perception task in DM&D.

Motivation for the current replication

DM&D's study provided an important contribution to the field, as it sought to better understand how inhibitory control might relate to perception and production in L2 speech. Critically, it demonstrated relationships between language abilities and general cognition, which provides support for usage-based accounts of L2 learning (Ellis, 2006; MacWhinney, 2008). To probe the reliability of the initial study's findings and to better understand which types of inhibitory control are most robustly related to L2 perception and production, the current study is a *close replication* (Porte & McManus, 2019, p. 73) of DM&D such that all major variables (e.g., participant characteristics, experimental tasks, materials, data analysis) remained unchanged except that inhibitory control was measured by two additional tasks: a Simon task and a Stroop task. Because the authors

shared materials and tasks on IRIS (https://www.iris-database.org/details/xSmdt-ui4Xg), replication using the original tasks and materials was feasible.

Research questions and replication assessment

DM&D (2016) did not explicitly state research questions (RQs), but rather identified that their aim was "to examine the relationship between the strength of L2 learners' inhibitory control and their accuracy in perceiving and producing L2 segments" (p. 745). The current study included three research questions with RQ1 intended to represent DM&D's initial question and RQ2 and RQ3 representing the variable modification.

- RQ1: To what extent is inhibitory control, as measured via a retrieval-induced inhibition task, related to L2 learners' accuracy in perception and production?
- RQ2: To what extent is inhibitory control, as measured via a Simon task, related to L2 learners' accuracy in perception and production?
- RQ3: To what extent is inhibitory control, as measured via a Stroop task, related to L2 learners' accuracy in perception and production?

Interpreting similarities and differences between the initial study and the current replication was done following the recommendations of Porte & McManus (2019): Means and standard deviations from the descriptive statistics of the two studies were used to compute Hedge's g effect sizes and corresponding confidence intervals (CIs) (see also McManus & Marsden, 2018). To evaluate between-study differences, the field-specific guidelines of Plonsky & Oswald (2014) were used such that Hedge's $g < .40$ and corresponding CIs passing through zero were taken to indicate negligible differences (see also McManus & Liu, 2022). For the main inferential tests (correlation and regression), directionality (positive vs. negative) and magnitude of the effect size were considered.

Method

All materials, experimental and coding protocols, data, and analysis code are publicly available at https://osf.io/fxzvj/, and the preregistration is available at https://osf.io/w4gj2. Table 1 summarizes the tasks used in the initial study and current replication.

Table 1. Summary of tasks for DM&D and the current replication

Materials	[^]Background questionnaire [a]X_Lex (proficiency)
Perception task Production task Cognitive tasks	[^]Speeded ABX categorization [^]Delayed sentence repetition [^]Retrieval–induced inhibition [*]Simon [*]Stroop

[^]Original tasks/materials provided by DM&D on IRIS.
[a]Original task not on IRIS—software used in the current replication is provided on IRIS/OSF.
[*]Tasks added for replication.

Participants and overall design

A power analysis using G*Power (Faul et al., 2009) based on the R^2 values from the hierarchical multiple regression in DM&D indicated a target sample size of 60 participants for the current replication (see Supplementary Materials, S1).[1] Seventy L1 English learners of L2 Spanish recruited from classes at the University of Pittsburgh participated in the experiment. Following DM&D, participants were excluded if they were early simultaneous bilinguals or their native speaker (NS) status was unclear ($n= 5$); they indicated having a speech/hearing pathology ($n = 1$); they scored below two standard deviations (SDs) of the group mean on the control condition in the perception task ($n = 3$); or they exhibited an extreme value in the retrieval-induced inhibition task ($n = 1$). An additional two participants were excluded because of missing production data ($n = 2$). Table 2 provides a comparison of the demographic characteristics between the participants[2] in DM&D and the final 58 L2 Spanish learners in the current replication.

Participant demographic information was gathered using the same language background questionnaire and motivation survey from the initial study. The motivation survey included nine statements related to language learning and use which participants rated on a 9-point Likert scale such that higher values correspond to greater motivation. To gauge L2 use, participants were asked to indicate how much (e.g., 0%, 1%–25%, 26%–50%) they used the L2 in certain contexts (e.g., texting, reading books). The percentage categories were converted to scores between 0 and 4 (e.g., 0% = 0, 76%–100% = 4) and an average score was calculated such that higher values correspond to greater use. As in the initial study, L2 proficiency was measured using X_Lex (Meara, 2005), a test of receptive vocabulary, as well as through self-report ratings (1–5) of how well participants estimated they could speak, understand, read, and write in Spanish (with higher values corresponding to greater proficiency). As indicated in Table 2, the

Table 2. Participant demographic comparison using means (SDs) of DM&D and the current replication

	DM&D	Current study	Effect size
	(*n*=18)	(*n*=58)	Hedge's *g*
	Mean (SD)	Mean (SD)	*g* [95% CI]
Age	19.50 (0.70)	20.07 (1.27)	−0.48 [−1.02, 0.05]
Motivation English (1–9)	7.35 (0.91)	7.09 (0.67)	0.35 [−0.18, 0.88]
Current L2 use (max 36)	9.11 (7.06)	6.98 (2.72)	0.51 [−0.03, 1.04]
Self–evaluation (1–5)	3.91 (0.58)	3.80[3] (0.50)	0.21 [−0.32, 0.74]
Speaking	3.61 (0.97)	3.54 (0.71)	0.09 [−0.44, 0.62]
Listening/understanding	4.11 (0.67)	3.70 (0.74)	0.56 [0.02, 1.10]
Residence abroad (weeks)	5.93 (15.10)	1.50 (5.79)	0.50 [−0.04, 1.03]
Years of study	8.77 (2.94)	9.02 (3.01)	−0.08 [−0.61, 0.45]
Age of first L2 exposure	8.83 (4.09)	7.97 (4.30)	0.20 [−0.33, 0.73]
Age of first L2 use	10.20 (3.96)	10.86 (3.98)	−0.16 [−0.69, 0.37]
L2 proficiency (X_Lex adjusted)	2,522 (721)	2,491 (676)	0.04 [−0.48, 0.57]

[1]Darcy et al. recruited $n = 26$ L2 Spanish learners from classes at Indiana University, of whom $n = 19$ were included in the perception and production analyses. The inhibition analyses reported included $n = 34$ participants ($n = 18$ L2 Spanish and $n = 16$ L2 English).

[2]Darcy et al. participant demographic information is taken from the Supplementary Materials (S7), as those were the L2 Spanish participants included in the inhibition correlation and regression analyses.

[3]Due to a data collection error in the current replication, self-rated proficiency data were missing from $n= 2$ participants.

participant demographic characteristics between the current study and DM&D are numerically similar. Hedge's *g* effect sizes and 95% CIs mostly indicated negligible effects. The only exception was self-evaluation of listening with a small effect and lower CI close to zero (*g* = 0.56 [0.02,1.10]).

Materials and procedure

The current study used the same instruments (https://www.iris-database.org/details/xSmdt-ui4Xg) and followed the same procedure as DM&D but incorporated two additional inhibition tasks. As in DM&D, data collection followed a similar order for each participant: production task, cognitive tasks, perception task, and vocabulary task. Data collection sessions lasted approximately 70 minutes. This research was approved by the University of Pittsburgh institutional review board; participants were compensated[4] with a $20 Amazon gift card.

Perception

Perception was tested using the same speeded ABX categorization task, stimuli, and procedure from DM&D administered via DMDX. The experimental items included the Spanish vowel and consonant contrasts /e-ei̯/ and /d-ɾ/. Control items with segmental contrasts native to English were also included. All stimuli were presented in trisyllabic nonce words (e.g., [faˈneða] or [fəˈnidɪʃ]). The stimuli were recorded in English and Spanish (i.e., with appropriate phonetic realizations for the respective language) by two female bilingual speakers of Mexican Spanish and American English. An additional set of control segments was included that consisted of vowel and consonant contrasts common to both English and Spanish. Each of the eight subconditions (two native, two nonnative, four control) were tested with four pairs of words.

The English and Spanish stimuli were presented in two separate blocks counterbalanced across participants. Within each block, stimuli were randomized. During each trial of the ABX task, participants heard three stimuli (nonce words) and had to select whether the last word they heard (X) matched the word of the first (A) or second (B) stimulus. Accuracy (i.e., correctly identifying which stimulus X matched) and reaction time (RT) were recorded. If a participant did not provide a response within 2,500 ms, the next trial began automatically. Within trials, a different voice was used for the A/B versus X recordings, and the recordings of the first and second stimulus were physically different. The four pairs of words for each of the eight subconditions were presented in four combinations (ABA, ABB, BAA, BAB) for a total of 128 trials (64 in each language). Participants completed the task on a laptop while wearing headphones. The task began with eight practice items that included feedback to familiarize participants with the procedure. A break was offered between the two blocks. The task took approximately 20 minutes.

Production

Production was tested using the same delayed sentence repetition task from DM&D administered via Microsoft PowerPoint. The consonant and vowel targets were the same as those in the perception task: /e-ei̯/ and /d-ɾ/. The task included a total of 16 sentences (four pairs per contrast) consisting of a question-and-answer sequence.

---

[4]Remuneration information was not explicitly stated in the initial study.

Each item was presented via a series of two slides. On the first slide, participants heard the question and answer (recorded by the same female speakers who produced perception stimuli) and simultaneously saw the sentences in their written form. On the second slide, participants heard the question again and had to repeat aloud the answer they had previously heard. The written forms were not provided on the second slide. While completing the task, participants wore headphones and were seated in front of a computer in a sound-controlled booth. Their responses were recorded with a Shure SM58 microphone via a Focusrite Scarlett 4i4 interface on a Dell computer. Participants who hesitated or misremembered a target word were asked to repeat the trial. Initial instructions and a practice item were provided in the L1 English. The task took approximately 7 minutes.

Inhibitory Control

Three inhibitory control tasks were administered in the current replication: the task used in the initial study (retrieval-induced inhibition) and two additional tasks (Simon and Stroop). The retrieval-induced inhibition task was administered using the same task and procedure from DM&D. All three inhibitory control tasks were administered via E-prime 2.0.8.22 in participants' L1 English. Participants were counterbalanced regarding the order in which they completed the tasks.

Retrieval-Induced Inhibition Task. The retrieval-induced inhibition task was the same task used in DM&D adopted from Lev-Ari and Peperkamp (2013). The task included three phases: In the first phase, 18 words were presented to the participants who were instructed to memorize the words. The 18 words were comprised of six words (e.g., tiger, duck, elephant, cow, horse, snake) belonging to each of three different categories (i.e., animals, occupations, or vegetables). Each word was displayed along with its category (e.g., "animal – tiger"). In the second phase, participants were instructed to recall words they had learned. Specifically, they were asked to recall three words from two of the categories by seeing a prompt such as "animal – t" and being directed to type their response. In the third phase, participants were presented with a list of 34 words and asked to identify whether the word was one of the 18 they had been asked to memorize during phase one of the experiment.

   The purpose of phase 2 was to create three types of experimental items: (a) words that were practiced, (b) words that were inhibited because they belonged to a category that was practiced but were not practiced themselves, and (c) words that functioned as a control because they belonged to a category that was not practiced. The expectation is that those with greater inhibitory control would have lower activation levels for words of type (b) and therefore would have longer reaction times for these items in Phase 3 compared with words of type (a) or type (c). Participants were automatically assigned by E-prime to one of six experimental lists based on their participant number. The task took approximately 8 minutes.

Simon Task. The Simon task used in the current replication was adapted from Lev-Ari and Peperkamp (2014) and represents domain-general inhibition of a prepotent response. In the task, participants were instructed to identify whether a box was red or blue by pressing "q" or "p" on a keyboard. The boxes appeared on the left, center, or right side of the screen. Participants needed to identify the color of the box while ignoring its location. Each item in the task represented one of three conditions: (i) a congruent condition in which the location of the box matched the location of the

keyboard button press (e.g., a box shown on the right hand side of the screen requiring a right/"p" button press); (ii) an incongruent condition in which the location of the box was the opposite of the location of the keyboard button press (e.g., a box shown on the right hand side of the screen requiring a left/"q" button press); and (iii) a neutral condition in which the box was located in the center of the screen. Reaction times are expected to be faster in congruent versus incongruent conditions.

The task began with a practice block consisting of 12 items (two cycles of each of the three conditions for each of the two colored boxes, randomly presented). During the practice block, participants received feedback (i.e., "incorrect" was displayed for any incorrect response). After the practice block, participants completed two blocks of 42 experimental items (total of 84 items), which were the same in procedure as the practice block except that no feedback was provided. Within each block, items were randomized. Participants were counterbalanced as to whether the blue box was associated with the left or right-hand side of the screen. The task took approximately 4 minutes.

Stroop Task.   The Stroop task used in the current study was a manual version of the task adapted from Lev-Ari and Peperkamp (2014) and represents the inhibition of a prepotent response. The task included four conditions: (i) a congruent condition in which the word semantics and ink color matched; (ii) an incongruent condition in which the word semantics and ink color conflicted; (iii) a neutral condition in which the symbol string @@@@ was presented; and (iv) a reading condition in which the word was presented in gray ink. Four colors were used: blue, red, green, and yellow.

The task began with a training component in which participants practiced matching a color to its correct response key (i.e., "s" for yellow, "d" for green, "k" for blue, and "l" for red). This phase included 100 trials using the symbols @@@@ 25 times in each of the colors. Next, participants practiced the full task with 40 items. Finally, participants completed two blocks of 96 experimental items (total of 192 items). Within each block, items were randomized. Participants received feedback during the training and practice sessions but not during the two experimental blocks. The task took approximately 10 minutes.

Coding and analysis

The current study followed the same coding procedures as reported in DM&D. Hedge's *g* effect sizes and 95% CIs were calculated using the Effect Size Calculator (https://www.cem.org/effect-size-calculator). The remaining analyses were conducted in R (Version 4.2.3; R Core Team, 2022).

Perception

Following DM&D, only the results for the perception accuracy data are reported; however, RT information from the DMDX output was used to identify unanswered items (any RT greater than 2,500 ms; 77 items out of a possible 7,424 or 1.0%) and accuracy (negative RTs indicated incorrect responses). In line with DM&D, the results were screened for outliers (i.e., performance below two SDs from the mean on the control condition), which resulted in three participants being excluded. Accuracy was calculated as an error rate for each participant based on performance of the Spanish consonant and vowel items.

Production

The segmental analysis followed the same procedure as described in DM&D for both vowel and consonant targets.

Vowel Analysis. The vowel analysis compared formant movement between the monophthong and diphthong vowels with the prediction that learners would produce the Spanish vowels /e-ei̯/ similarly. Each vowel was segmented in Praat (Boersma & Weenink, 2023) independently by both the first author and a research assistant following a segmentation protocol (provided on OSF) to allow for comparison and checking. Using the segmented TextGrids and a Praat script modified from Brato (2016)[5] measurement points at 20%, 50%, and 80% were identified. To analyze formant movement, the average F0 (fundamental frequency), F1, and F2 were measured in Hertz over 10-ms windows centered at the 20%, 50%, and 80% locations. For normalization purposes, these frequency values were converted to Bark using the formula provided by Baker & Trofimovich (2005, p. 9): $B = 26.81/(1 + [1{,}960/F]) - 0.53$. Estimates of vowel position in terms of height and frontness were calculated using the Bark converted frequencies: height ($B1 - B0$) and frontness ($B2 - B1$). The overall formant movement was estimated: First, the Euclidean distances between the 20% and 50% measurement points and the 50% and 80% measurement points were calculated. These two distances were summed and used to represent the amount of formant movement within the vowel. Finally, following DM&D, a *z* score was computed for the learners' diphthong formant movement based on the means and SDs reported for the NSs ($M = 3.19$, $SD = 0.71$, p. 747).

Consonant Analysis. The L2 Spanish consonants /d-ɾ/ were binarily scored (accurate or inaccurate) by two native speaker raters based on auditory and visual acoustic characteristics (see protocol on OSF). Following DM&D, to be coded as accurate, /d/ needed to be realized as spirantized [ð] and /ɾ/ needed to be realized as a single-closure tap with short duration. With four tokens of each consonant, participants received a score of a maximum of eight. After training and group rating of two experimental files, the two raters separately coded the remaining files. They agreed on 93% of the ratings, and the remaining differences were discussed and resolved. Interrater reliability was acceptable and comparable to DM&D (replication $k = .92$, DM&D $k = .92$).

Inhibitory Control

Retrieval-Induced Inhibition. Following DM&D, participants' results were first checked to ensure that they recalled two or more items out of six in the practiced categories. All participants in the current replication passed this check. Next, median RTs were computed for each participant; comparisons were made among the item types; and an inhibitory control score was computed by dividing the median RT for inhibited items by the median RT for control (non-practiced) items. In this way, the greater the score is above 1, the stronger the inhibitory control.

Simon. A Simon effect is calculated by subtracting RTs on congruent trials from RTs on incongruent trials (Linck et al., 2012) for correct responses, and those with higher

_____
[5]The formant analysis exactly followed the procedure described by Darcy et al., but they did not use the Praat script.

inhibitory control therefore have lower Simon effects. Average RTs (for correct responses only) were calculated for each participant by trial type (i.e., congruent, neutral, or incongruent), and a Simon score was computed for each participant by subtracting the average RT on congruent trials from the average RT on incongruent trials.

Stroop. A Stroop effect is calculated by subtracting RTs on the neutral trials from RTs on the incongruent trials (Gass & Lee, 2011; Lev-Ari & Peperkamp, 2014) for correct responses, and those with higher inhibitory control therefore have lower Stroop effects. Average RTs (for correct responses only) were calculated for each participant by trial type (i.e., congruent, incongruent, neutral, or read), and a Stroop score was computed for each participant by subtracting the average RT on neutral trials from the average RT on incongruent trials.

Relationship between inhibitory control and phonological processing

In line with DM&D, the following variables were included in partial correlation analyses: Perception scores were the error rates from the ABX perception task; consonant production was the number of accurate productions (out of eight); and vowel production was a *z* score of diphthong formant movement computed using Spanish NS means/SDs from DM&D. Following DM&D and as planned in the study preregistration, two main analyses were conducted and proficiency (X_Lex score) was included as a covariate: (i) partial correlation analyses between inhibitory control and the phonological scores, and (ii) hierarchical regression analysis using inhibitory control as a predictor of vowel perception.

The test assumptions of the Pearson partial correlation analyses were checked (i.e., linear relationships among variables, normal distribution, and lack of outliers; see Supplementary Materials, S2). Assumptions were violated (i.e., the phonological variables did not appear to be normally distributed nor did transformations of the variables improve distribution). Therefore, non-parametric Spearman partial correlation analyses were conducted using the *PResiduals* package *partial_Spearman*.

Finally, test assumptions for the hierarchical multiple regression analysis were checked (e.g., linear relationships among variables, homoscedasticity of residuals, lack of influential points; see RegressionAnalysis.R code). Multiple assumptions were violated, including the existence of influential points. Removal of the points and transformation did not result in model improvements. Therefore, the results of the regression analysis are interpreted cautiously by visual inspection of the residuals plot.

Results

Perception

Table 3 reports the descriptive statistics for the vowel and consonant perception error rates in DM&D and the current replication.

As indicated in Table 3, the vowel perception error rates in the current study and DM&D are numerically similar, with average vowel and consonant perception accuracy being 80.9% and 76.0%, respectively, in DM&D, and 80.8% and 83.2% in the current replication. Comparing the SDs and CIs, we see larger SDs in the initial study compared with the current replication, but at the same time the CI widths are relatively

Table 3. Comparison of the descriptive statistics for vowel and consonant perception error rates between DM&D and the current replication

	DM&D	Current study	Effect size
	(*n* = 19)	(*n* = 58)	Hedge's *g*
Vowel			
Mean (SD)	19.1 (20.0)	19.2 (13.2)	−0.01 [−0.52, 0.51]
Range	0.0–75.0	0.0–63.6	
95% CI	[15.5, 22.7]	[15.8, 22.7]	
Consonant			
Mean (SD)	24.0 (23.0)	16.8 (14.0)	0.43 [−0.09, 0.95]
Range	0.0–75.0	0.0–53.3	
95% CI	[20.4, 27.6]	[13.1, 20.4]	

Table 4. Comparison of the means (SDs) for vowel (formant movement) and consonant (target-like rating) production accuracy between DM&D and the current replication

	DM&D	Current study	Effect size
	(*n* = 19)	(*n* = 58)	Hedge's *g*
/ei̯/ formant	1.18 (0.55)	1.58 (0.60)	−0.67 [−1.20, −0.14]
	n/a	[1.42, 1.74]	
/e/ formant	0.92 (0.36)	1.26 (0.38)	−0.90 [−1.43, −0.36]
	n/a	[1.16, 1.36]	
/ei̯/ ~ /e/	*t*(18) = 1.92, *p* = 0.71	*t*(57) = 3.68, *p* <. 001	
	d = 0.56, [−0.10, 1.20]	*d* = 0.64, [0.26, 1.01]	
/ei̯/ *z* score	−2.54 (0.78)[a]	−2.27 (0.85)	−0.32 [−0.85, 0.21]
Consonant acc. (max. 8)	4.09 (2.45)[a]	4.50 (2.31)	−0.17 [−0.70, 0.36]
	n/a	[3.89, 5.11]	

[a]Value from p. 761 L2 speakers (*n* = 18) included in the inhibitory control analyses.

similar in both studies (about seven for both consonants and vowels). This may indicate potential outliers in the initial study. The Hedge's *g* effect sizes and 95% CIs for these mean difference comparisons indicated negligible effects (with CIs passing through/including zero) for both vowel (*g* = −0.01 [−0.52, 0.51]) and consonant (*g* = 0.43 [−0.09, 0.95]) perception error rates, suggesting negligible differences between the performances of the two groups of participants.

Production

Table 4 reports the descriptive statistics for the vowel and consonant production accuracy scores from DM&D and the current replication. L2 learners in DM&D did not demonstrate clear differences in formant movement between the monophthong and diphthong vowels: *M* = 0.92 (0.36) compared with *M* = 1.18 (0.55), *g* = 0.55 [−0.10, 1.20]; however, L2 learners in the current replication did show differences with a small effect size, *M* = 1.58 (0.60) compared with *M* = 1.26 (0.38), *g* = 0.63 [0.26, 1.01]. Nevertheless, as in DM&D, Spanish L2 learners produced the diphthong /ei̯/ with much less formant movement than NSs: *M* = 1.58 (0.60) compared with *M* = 3.19 (0.71), *g* = 2.61 [1.65, 3.56]. It was also the case that the *z* score values computed
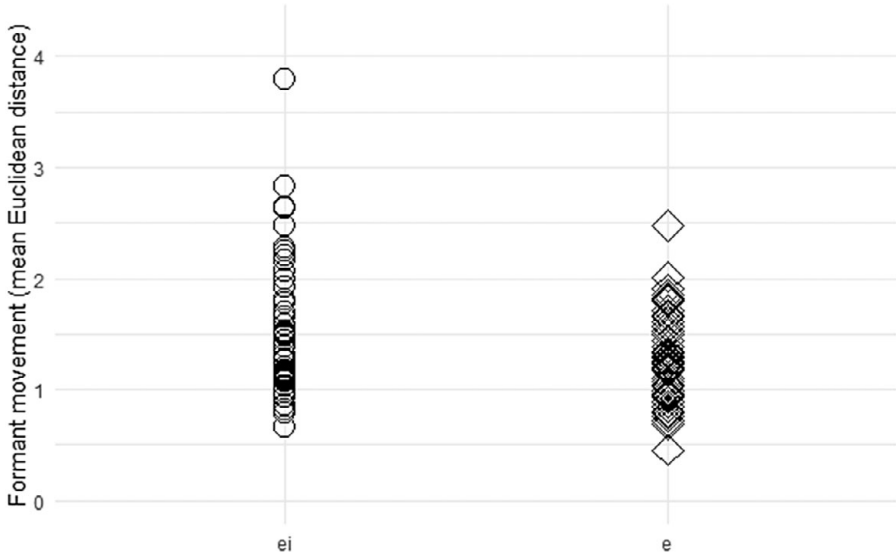
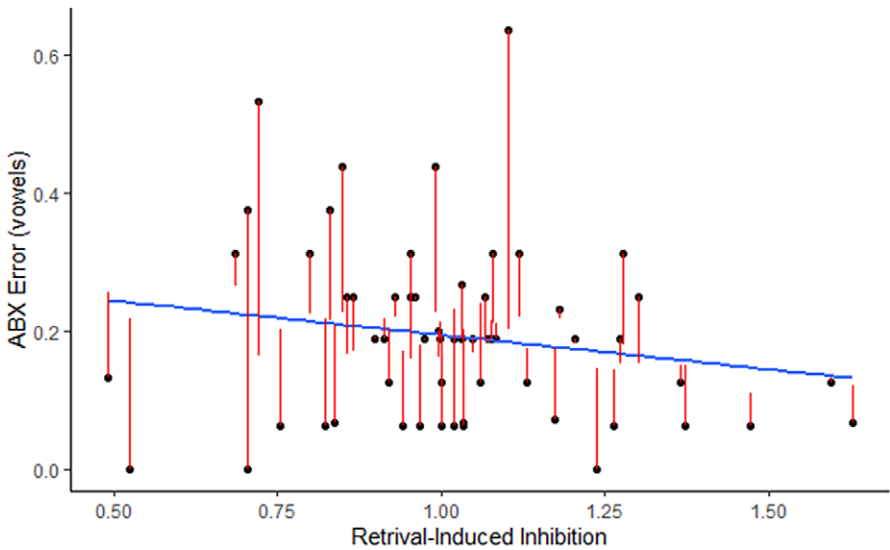**Figure 1.** L2 learners' formant movement for diphthongs and monophthongs.

Figure 2. Scatterplot of vowel perception (ABX error rate) and retrieval-induced inhibition with regression line and residual error.

for the L2 learners' diphthong formant movement were comparable between the studies: $M = -2.54$ (0.78) compared with $M = -2.27$ (0.85), $g = -0.32$ [-0.85, 0.21]. Figure 1 provides a visual comparison of the formant movement in the vowels produced by L2 learners in the current study and is parallel to Figure 2 reported in DM&D (p. 758).

Table 5. Comparison of the descriptive statistics for the retrieval-induced inhibition task between DM&D and the current replication

	DM&D L2 Spanish (*n* = 36[b])	Current study L2 Spanish (*n* = 58)
Practiced items	807 ms	768 ms (254) [701, 835]
Inhibited items	985 ms	850 ms (208) [796, 905]
Control (non–practiced) items	933 ms	869 ms (297) [791, 947]
Practice–Inhibited	*t*(35) = –3.05, *p* =. 002	*t*(57) = –2.57, *p* =. 013 *d* = –0.35 [–0.72, 0.02]
Control–Inhibited	*t*(35) = –0.60, *p* > .10	*t*(57) = 0.59, *p* =. 559 *d* = 0.07 [–0.29, 0.44]

[b]Values reported include both L2 speaker groups from DM&D, as results for the L2 Spanish only group were not reported (p. 760).

For the /d-ɾ/ consonant contrast, participants in the current replication scored similarly (*M* = 4.50 [2.31] out of eight) in comparison to participants in DM&D (*M* = 4.09 [2.45]), with effect-size 95% CIs passing through zero (*g* = –0.17 [–0.70, 0.36]).

Inhibitory control

Retrieval-Induced Inhibition

As indicated in Table 5, results in the current study patterned similarly to those of DM&D when comparing retrieval-induced inhibition task performance: (i) Median RTs for the practice items were shortest of the three item types, and (ii) although a significant difference was found between practiced versus inhibited items, no significant difference was found between control versus inhibited items. For the inhibitory control scores, as indicated in Table 8, participants in the current replication scored similarly (1.02 [0.23]) to participants in DM&D (1.01 [0.11]), with effect size CIs passing through zero (*g* = –0.05 [–0.58, 0.48]).

Simon. Mean RTs and SDs (correct responses only, 97% of total responses) for each trial type on the Simon task are presented in Table 6. Results of the Simon task indicated longer RTs for incongruent conditions (*M* = 518, *SD* = 180) compared with congruent conditions (*M* = 479, *SD* = 147) with non-overlapping CIs.

Stroop. Mean RTs and SDs (correct responses only, 96% of total responses) for each trial type on the Stroop task are presented in Table 7. Results of the Stroop task indicated longer RTs for incongruent conditions (*M* = 1,188, *SD* = 648) compared with neutral conditions (*M* = 857, *SD* = 414) with non-overlapping CIs.

Table 6. Means (and SDs) of RTs for each trial type on the Simon task

	Mean (ms)	SD	95% CI
Congruent	479	147	[472, 486]
Neutral	488	191	[479, 497]
Incongruent	518	180	[509, 527]
Simon effect	39	45	

Table 7. Means (and SDs) of RTs for each trial type on the Stroop task

	Mean (ms)	SD	95% CI
Congruent	850	424	[834, 866]
Incongruent	1,188	648	[1,163, 1,213]
Neutral	857	414	[841, 873]
Read	975	480	[957, 994]
Stroop effect	331	206	

As planned in the study preregistration, Simon and Stroop effect scores were tested for normality. Because neither Simon nor Stroop scores were normally distributed (see Supplementary Materials, S3), trial-level RTs were log-transformed and median log-RTs were computed, as recommended by Linck et al. (2012). For both Simon and Stroop effects, this resulted in scores that were normally distributed; therefore, the log-RTs were used in partial correlation and regression analyses.

Relationship between inhibitory control and phonological processing

Table 8 reports the means and SDs of the target phonological variables and three measures of inhibitory control included in the analysis. As a reminder, DM&D did not conduct an analysis for consonant perception because the L2 English group performed at ceiling on the perception task; thus, Table 8 contains "n/a" values for consonant (ABX) error rates.

For ease of interpretation, Table 9 summarizes the expected directionality of the relationships for the partial correlation analyses among the phonological variables and the tests of inhibitory control used in the current study. For example, vowel perception is operationalized as error rate on the ABX task, so a higher score corresponds to lower accuracy. For the retrieval-induced inhibition score, higher inhibitory control is associated with longer retrieval RTs, so a higher score corresponds to higher inhibitory

Table 8. Means and SDs of target cognitive and phonological variables in DM&D and the current replication

	DM&D L2 Spanish ($n = 18$)	DM&D All learners ($n = 34$)	Current study L2 Spanish ($n = 58$)
Retrieval–induced inhibition	1.01 (0.11)	1.03 (0.21)	1.02 (0.23)
ABX error (vowels)	0.20 (0.14)	0.24 (0.14)	19.2 (13.2)
ABX error (consonants)	n/a	n/a	16.8 (14.0)
Vowel production z score	−2.54 (0.78)	−3.08 (0.86)	−2.27 (0.85)
Consonant production (max. 8)	4.09 (2.45)	5.41 (2.42)	4.50 (2.31)

Table 9. Expected directionality of the relationships among the phonological and inhibition scores

	Retrieval-induced	Simon	Stroop
Vowel perception	Inverse/negative*	Direct/positive	Direct/positive*
Consonant perception	Inverse/negative*	Direct/positive*	Direct/positive*
Vowel production	Direct/positive*	Inverse/negative	Inverse/negative
Consonant production	Direct/positive*	Inverse/negative*	Inverse/negative

*Expected directionality of the relationship found in the current study.

Table 10. Partial correlations between phonological variables and retrieval-induced inhibition in DM&D and the current replication

	DM&D All learners (*n* = 33)	Current study L2 Spanish (*n* = 58)
ABX error (vowels)	*r* = −0.42**	r_s = −0.13 [−0.41, 0.16], *p* =. 379
ABX error (consonants)	n/a	r_s = −0.11 [−0.37, 0.17], *p* =. 451
Vowel production *z* score	*r* = −0.04	r_s = 0.09 [−0.19, 0.35], *p* =. 529
Consonant production (max. 8)	*r* = 0.34*	r_s = 0.07 [−0.20, 0.33], *p* =. 610

Note: * *p* < .05, ** *p* < .01 as reported in DM&D.

control. Thus, an inverse or negative relationship is predicted between vowel perception scores and retrieval-induced inhibition scores if it is the case that stronger inhibitory control is related to more accurate vowel perception abilities. In contrast, for the Simon task, higher inhibitory control is associated with a lower Simon score, so a higher score corresponds to lower inhibitory control. Thus, a direct/positive relationship is predicted between vowel perception scores and Simon scores if it is the case that stronger inhibitory control is related to more accurate vowel perception abilities. The asterisks in Table 9 indicate when the expected directionality of the relationship was indeed found in the current replication study. These results are discussed in more detail in the Discussion.

Table 10 reports the results of the partial correlation analyses (Pearson for DM&D, Spearman for the current study) for comparison. DM&D reported statistically significant relationships with medium effect (*r* = −0.42 and 0.34) between inhibitory control and vowel perception and inhibitory control and consonant production, such that participants with higher inhibitory control were more accurate in vowel perception and consonant production. No clear relationship was reported between vowel production and inhibitory control. In the current replication, the directionality of the relationships is as expected such that higher inhibitory control corresponded to higher perception and production accuracy. Nevertheless, no statistically significant relationships were found, and 95% CIs for the effect sizes (all considered less than small) passed through zero, indicating negligible effects.

Table 11 reports the results of the Spearman partial correlation analyses with the two additional measures of inhibitory control included in the current analysis.

As with the results for retrieval-induced inhibition, no statistically significant relationships were found, and 95% CIs for the effect sizes (all considered less than small) passed through zero, indicating negligible effects. If the directionality of the relationships is considered, for the Simon task, we find the predicted relationship for the consonant targets (higher inhibitory control corresponding to more accurate consonant perception and production); however, this is not the case for the vowel targets such that higher inhibitory

Table 11. Partial correlations (Spearman) between phonological variables in the Simon and Stroop tasks

	Simon	Stroop
ABX error (vowels)	−0.12 [−0.37, 0.16], *p* =.407	0.18 [−0.08, 0.42], *p* =. 170
ABX error (consonants)	0.07 [−0.17, 0.30], *p* =. 588	0.23 [−0.02, 0.45], *p* =. 069
Vowel production *z* score	−0.18 [−0.44, 0.11], *p* =. 221	0.13 [−0.13, 0.38], *p* =. 331
Consonant production (max. 8)	0.01 [−0.24, 0.27], *p* =. 924	0.10 [−0.16, 0.34], *p* =. 470

Table 12. Hierarchical regression results with retrieval-induced inhibition predicting vowel perception error rates while controlling for proficiency

Variable	Predictor	R^2	B	95% CI	t	p
DM&D	X_Lex	0.00	0.02	0.00, 0.00	0.13	.901
ABX	Inhibition	0.18	−0.42	−0.49, −0.05	−2.51	.018
Replication	X_Lex	0.02	−0.00	0.00, 0.00	−1.54	0.129
ABX	Inhibition	0.03	−0.09	−0.24, 0.07	−1.13	0.264

control was associated with less accurate vowel perception and vowel production. For the Stroop task, a third pattern emerged such that we found the predicted relationships for the perception targets (both consonant and vowel) but not the production targets.

Finally, results from the hierarchical multiple regression analysis are reported. Table 12 reports the results of the hierarchical multiple regression analysis from DM&D along with the values from the current replication (although recall that test assumptions were violated). DM&D reported that inhibitory control was a statistically significant predictor of vowel perception accuracy ($p = .018$) with a small effect[6] ($R^2 = 0.18$) once proficiency was controlled. In the current replication, retrieval-induced inhibition was not a statistically significant predictor of vowel perception accuracy ($p = .264$, $R^2 = 0.03$). Figure 2 graphically illustrates this relationship via a scatterplot of ABX error rate (vowels) and retrieval-induced inhibition with the fitted regression line in blue and the residual errors (difference between the observed and predicted value) in red.

Discussion

The current study was a close replication of Darcy et al. (2016), which examined the relationship between inhibitory control and L2 speech perception and production and unexpectedly reported differential relationships between inhibitory control and perception and production. Specifically, DM&D found that learners with higher inhibitory control were more accurate in vowel perception and consonant production. No clear relationship was found between vowel production and inhibitory control, and consonant perception was not able to be tested. Results of the current replication for the partial correlation analyses did not indicate statistically significant relationships, and effect sizes were small with 95% CIs crossing through zero. Assumptions were not met for the hierarchical multiple regression analysis. At the same time, a comparison of the descriptive statistics (means and SDs) for the tests of phonological processing and retrieval-induced inhibition between DM&D and the current replication indicated negligible differences between the two participant samples in terms of effect sizes and 95% CIs.

The underlying theoretical question at the core of DM&D and the current replication regards the extent to which L2 phonological skills are related to general cognitive abilities, specifically inhibitory control. The logic behind the hypothesized relationship is that those with greater inhibitory control might be better at suppressing their L1 during the processing of L2 acoustic-phonetic input, ultimately resulting in more accurate segmental categories. Having these more accurate categories would, in turn, result in more accurate perception and production of segments during language use.

---

[6]Interpreted using the field-specific recommendations from Plonsky and Ghanbar (2018).

Put another way, variability in outcomes in L2 pronunciation could be the result of individual differences in inhibitory control.

Bringing together the findings from DM&D and the current replication, no strong, clear, or consistent relationship emerges between inhibitory control and L2 perception/production skills. There are multiple explanations for why there were discrepancies between the results of the two studies. The following possibilities are discussed in more detail here: (i) There is no, or only a weak, relationship between inhibition and L2 phonological skills; (ii) the type of inhibition tested or the specific tests used to measure inhibition were not appropriate to capture individual differences in cognitive abilities; and/or (iii) there were differences in study design features (e.g., participant samples, methodological decisions, time frame, context) that affected outcomes.

When considering the combined results of DM&D and the current replication, an important possibility to consider is that despite inhibitory control (both retrieval-induced and prepotent response inhibition) being demonstrated to relate to language processing in sentence comprehension and lexical access studies (Costa & Santesteban, 2004; Filippi et al., 2012), its relationship to phonological processing may be null or weak at best. Even if inhibitory control affects L2 pronunciation skills, it could be that it plays a more critical role in acquisition stages or perceptual category formation; therefore, perception measures that reflect those processes might be more likely to show relationships as opposed to those in the current study, which reflect more end-state performance. In fact, an intriguing body of research is providing mounting evidence for the role of domain-general auditory processing in second language learning (e.g., Saito et al., 2021, 2022) such that "the ability to precisely encode auditory input [e.g., information about frequency, duration, amplitude] may be a bottleneck for the establishment of knowledge about segmental and suprasegmental linguistic categories" (Zheng et al., 2022, p. 480). Additional evidence links auditory processing to the establishment of higher executive functions via the temporal and sequential patterns present in sound signals and supports frameworks such as those suggested by the Auditory Scaffolding Hypothesis (Conway et al., 2009). As such, to better understand the relationship between inhibitory control and L2 phonological skills, a useful future replication study could test learners' auditory processing skills and incorporate those scores in statistical models.

Another important consideration is that the specific tests used to measure inhibition might not have been appropriate to capture individual differences in cognitive abilities. Regarding the retrieval-induced inhibition task, a finding that is consistent across both studies was that inhibited items were responded to slower than practiced items. Neither study indicated significant differences between the inhibited and control items (recall that the inhibitory control score was computed by dividing the median RT for inhibited items by the median RT for control items). In other words, neither study found the expected differences between RTs on control versus inhibited items, which may be an indication that the task is not sensitive enough for this population. Future work examining the role of inhibitory control with language learning and should more thoroughly consider performance on the tasks used to measure inhibition. For instance, it is necessary to consider whether there is sufficient variability in inhibition scores. The same is true for the Stroop and Simon tasks. At first it might seem somewhat counterintuitive that highly robust and reliable tasks like the Stroop task would be problematic for investigations of individual differences using correlational analyses. Yet, Hedge

et al. (2018) provided convincing evidence that such tasks are robust *because* they lack between-participant variability, which is necessary for ranking individuals for correlational studies like the current replication. This reduced reliability also results in reduced power for analysis, meaning that the sample sizes needed to detect relationships should likely be two to three times larger than those calculated without taking task reliability into account (Hedge et al., 2018). For the current study that would mean about 180 participants. This might not be feasible in all cases, but the open sharing of data in the field would mean that researchers could combine data sets over time.

The current replication primarily explored whether L2 phonological processing would be related to tests of inhibitory control designed to measure prepotent response inhibition. Specifically, it sought to explore whether Simon/Stroop tasks would show a stronger relationship to production measures, given the possibility that production might additionally require the inhibition of a dominant motor response (i.e., the articulation of an L1 segment). Similar to the findings for retrieval-induced inhibition, no significant relationships were found, and effect sizes were small with 95% CIs crossing through zero. When considering the directionality of the results, different patterns emerged in comparison to retrieval-induced inhibition: Although the expected directionality of the relationships was present for perception and production of consonants and vowels for the retrieval-induced inhibition task, this was only the case for consonant measures (both perception and production) for the Simon task and perception measures (both consonant and vowel) for the Stroop task. In their study, DM&D summarized their findings by stating, "In sum, the relationship between inhibition and perception appeared stronger than the relationship between inhibition and production" (p. 764). One interesting complication, however, is that the pattern of the strengths of the relationships was not identical for vowel and consonant targets. A stronger relationship was noted for perception, but this was based only on the vowel perception results, as consonant perception data were not able to be analyzed in the initial study. Thus, it is likely that relationships between L2 perception and production and inhibitory control might differ when the target feature is a vowel or consonant. It is not immediately clear why this might be the case. One possibility is that it is not consonant versus vowel, per se, which is important, but rather the relative difficulty of perceiving and/or producing the target segment. Another possible explanation could be methodological if one considers the fact that vowel and consonant accuracy were measured differently on the production task in the studies. As acknowledged by DM&D, the formant analysis might allow for more individual variation than the consonant analysis, which was scored binarily and out of 8. At the same time, error rates for consonant and vowel perception would have resulted in comparable score types.

A final consideration relates to differences in study design features that might have affected outcomes. One obvious difference between the studies is that the sample sizes differed (n = 18, 19, 34 in DM&D and n = 58 in the current replication). The potentially underpowered initial study may have increased the risk of type I or type II errors (Loewen & Hui, 2021). One benefit of a larger sample is that it most likely results in narrower confidence intervals for effect sizes. As the field continues to improve on reporting practices, it is critical for researchers to report effect-size CIs, to allow for comparison in future replications (see Tables 10 and 11). Despite the lack of finding similar effects for the correlation analyses, a comparison of the descriptive statistics indicated comparability between the two studies regarding the outcomes of the phonological measures and the retrieval-induced inhibition task. One difference

between the current replication and the initial study is that the correlation analyses in the initial study combined phonological (and inhibition) scores from two different L2 groups. In the current study, the descriptive statistics for the phonological scores were comparable to those of the L2 Spanish group in the initial study and the inhibition scores of the combined group for the correlation analysis. Nevertheless, the fact that the initial study's correlation analyses included data from an L2 English group (whose phonological descriptives may have differed from the current sample) may have impacted outcomes.

In sum, the combined evidence from DM&D and the current replication does not provide consistent evidence of a strong relationship between inhibitory control and L2 perception/production skills. It could be that this relationship is mediated by general auditory processing skills and/or that our tests used to measure inhibition are not appropriate to capture individual differences. Clearly, we have much more to investigate to grow our understanding of the links between cognitive mechanisms and L2 phonological processing. For all potential future replications of this line of research, it is hoped that the open sharing of materials, data, and analysis code from the current study are useful.

References

Baker, W., & Trofimovich, P. (2005). Interaction of native- and second-language vowel system(s) in early and late bilinguals. *Language and Speech*, *48*, 1–27. https://doi.org/10.1177/00238309050480010101

Bearden, R. L., Asgari, S., Sobel, K. V., & Scoles, M. T. (2021). Manual responses are verbally mediated in Stroop identification tasks. *Psi Chi Journal of Psychological Research*, *26*, 121–129. https://doi.org/10.24839/2325-7342.JN26.2.121

Boersma, P., & Weenink, D. (2023). *Praat: Doing phonetics by computer.* Version 6.3.02. http://www.praat.org/

Botezatu, M. R., Guo, T., Kroll, J. F., Peterson, S., & Garcia, D. L. (2022). Sources of variation in second and native language speaking proficiency among college-aged second language learners. *Studies in Second Language Acquisition*, *44*, 305–330. https://doi.org/10.1017/S0272263121000188

Brato, T. (2016). TB_Track Vowels 2.1 [Praat Script]. https://www.uni-regensburg.de/language-literature-culture/english-linguistics/staff/brato/praat-scripts/index.html

Conway, C. M., Pisoni, D. B., & Kronenberger, W. G. (2009). The importance of sound for cognitive sequencing abilities: The auditory scaffolding hypothesis. *Current Directions in Psychological Science*, *18*, 275–279. https://doi.org/10.1111/j.1467-8721.2009.01651.x

Costa, A., & Santesteban, M. (2004). Lexical access in bilingual speech production: Evidence from language switching in highly proficient bilinguals and L2 learners. *Journal of Memory and Language*, *50*, 491–511. https://doi.org/10.1016/j.jml.2004.02.002

Darcy, I., Mora, J. C., & Daidone, D. (2016). The role of inhibitory control in second language phonological processing. *Language Learning*, *66*, 741–773. https://doi.org/10.1111/lang.12161

Dijkstra, T., & Van Heuven, W. J. (2002). The architecture of the bilingual word recognition system: From identification to decision. *Bilingualism: Language and Cognition*, *5*, 175–197. https://doi.org/10.1017/S1366728902003012

Ellis, N. C. (2006). Selective attention and transfer phenomena in L2 acquisition: Contingency, cue competition, salience, interference, overshadowing, blocking, and perceptual learning. *Applied Linguistics*, *27*, 164–194.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*, 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

Filippi, R., Leech, R., Thomas, M. S., Green, D. W., & Dick, F. (2012). A bilingual advantage in controlling language interference during sentence comprehension. *Bilingualism: Language and Cognition*, *15*, 858–872. https://doi.org/10.1017/S1366728911000708

Friedman, N. P., & Miyake, A. (2004). The relations among inhibition and interference control functions: A latent-variable analysis. *Journal of Experimental Psychology: General*, *133*, 101–135. https://doi.org/10.1037/0096-3445.133.1.101

Gass, S., & Lee, J. (2011). Working memory capacity, inhibitory control, and proficiency in a second language. In M. S. Schmid & W. Lowie (Eds.), *Modeling bilingualism: From structure to chaos* (pp. 59–84). John Benjimins.

Green, D. W. (1998). Mental control of the bilingual lexico-semantic system. *Bilingualism: Language and Cognition*, *1*, 67–81. https://doi.org/10.1017/S1366728998000133

Green, D. W., & Abutalebi, J. (2013). Language control in bilinguals: The adaptive control hypothesis. *Journal of Cognitive Psychology*, *25*, 515–530. https://doi.org/10.1080/20445911.2013.796377

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods*, *50*, 1166–1186. https://doi.org/10.3758/s13428-017-0935-931

Kim, D., Clayards, M., & Kong, E. J. (2020). Individual differences in perceptual adaptation to unfamiliar phonetic categories. *Journal of Phonetics*, *81*, 100984. https://doi.org/10.1016/j.wocn.2020.100984

Lev-Ari, S., & Peperkamp, S. (2013). Low inhibitory skill leads to non-native perception and production in bilinguals' native language. *Journal of Phonetics*, *41*, 320–331. https://doi.org/10.1016/j.wocn.2013.06.002

Lev-Ari, S., & Peperkamp, S. (2014). The influence of inhibitory skill on phonological representations in production and perception☆. *Journal of Phonetics*, *47*, 36–46. https://doi.org/10.1016/j.wocn.2014.09.001

Linck, J. A., Schwieter, J. W., & Sunderman, G. (2012). Inhibitory control predicts language switching performance in trilingual speech production. *Bilingualism: Language and Cognition*, *15*, 651–662. https://doi.org/10.1017/S136672891100054X

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

Luque, A., & Morgan-Short, K. (2021). The relationship between cognitive control and second language proficiency. *Journal of Neurolinguistics*, *57*. https://doi.org/10.1016/j.jneuroling.2020.100956

MacWhinney, B. (2008). A unified model. In P. Robinson & N. C. Ellis (Eds.), *Handbook of cognitive linguistics and second language acquisition* (pp. 341–371). Routledge.

McManus, K. (2021). *Crosslinguistic influence and second language learning*. Routledge.

McManus, K., & Liu, Y. (2022). Using elicited imitation to measure global oral proficiency in SLA research: A close replication study. *Language Teaching*, *55*, 116–135. https://doi.org/10.1017/S026144482000021X

McManus, K., & Marsden, E. (2018). Online and offline effects of L1 practice in L2 grammar learning: A partial replication. *Studies in Second Language Acquisition*, *40*, 459–475. https://doi.org/10.1017/S0272263117000171

Meara, P. M. (2005). *X_Lex: The Swansea Vocabulary Levels Test (Version 2.05)* [Computer software]. Swansea, UK: Lognostics.

Plonsky, L., & Ghanbar, H. (2018). Multiple regression in L2 research: A methodological synthesis and guide to interpreting R^2 values. *Modern Language Journal*, *102*, 713–731. https://doi.org/10.1111/modl.12509

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

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

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Saito, K., Suzukida, Y., Tran, M., & Tierney, A. (2021). Domain-general auditory processing partially explains second language speech learning in classroom settings: A review and generalization study. *Language Learning*, *71*, 669–715. https://doi.org/10.1111/lang.12447

Saito, K., Sun, H., Kachlicka, M., Alayo, J. R. C., Nakata, T., & Tierney, A. (2022). Domain-general auditory processing explains multiple dimensions of L2 acquisition in adulthood. *Studies in Second Language Acquisition*, *44*, 57–86. https://doi.org/10.1017/S0272263120000467

Simon, J., & Rudell, A. (1967). Auditory S–R compatibility: The effect of an irrelevant cue on information processing. *Journal of Applied Psychology*, *51*, 300–304. https://doi.org/10.1037/h0020586

Stroop, J. R. (1935). Studies of serial verbal reactions. *Journal of Experimental Psychology*, *18*, 643–662. https://doi.org/10.1037/h0054651

Zheng, C., Saito, K., & Tierney, A. (2022). Successful second language pronunciation learning is linked to domain-general auditory processing rather than music aptitude. *Second Language Research*, *38*, 477–497. https://doi.org/10.1177/0267658320978493