Language Teaching (2024), **57**, 419–432
doi:10.1017/S0261444823000083

CAMBRIDGE UNIVERSITY PRESS

REPLICATION RESEARCH

Optimizing second language pronunciation instruction: Replications of Martin and Sippel (2021), Olson and Offerman (2021), and Thomson (2012)

Charlie Nagle[1]* and Phil Hiver[2]

[1]The University of Texas at Austin, Austin, USA and [2]Florida State University, Tallahassee, USA
*Corresponding author. Email: cnagle@austin.utexas.edu

(Received 22 February 2023; accepted 1 March 2023)

Abstract

An important shift in language learning research is the understanding that pronunciation instruction is necessary to ensure learners' balanced development in pronunciation and second language (L2) speech. Research focusing on understanding what types of pronunciation instruction are effective and what makes them most effective has grown dramatically over the past decade. Given the methodological heterogeneity apparent in this body of literature, however, many questions remain to understand the specific effects of pronunciation training paradigms on learners' L2 development. In this article, we make the case that replication is a productive means of validating findings and assessing the strength of existing evidence in L2 pronunciation research. As prime targets for replication, we review three studies that offer different perspectives on how L2 pronunciation instruction can be optimized. We have chosen each of these studies because they significantly advance the field conceptually, have broken new ground empirically, and point to areas in which replication studies can have a large impact. We describe a series of close and/or approximate replications of each initial study that would add detail to existing knowledge and provide a more comprehensive understanding of how and why L2 pronunciation instruction is effective, for whom, and under what conditions.

1. Introduction

Second language (L2) researchers have always been concerned with the role instruction plays in L2 development. Norris and Ortega (2000) were among the first to address this question from a meta-analytic framework, finding that L2 instruction, especially explicit instruction, on average leads to large gains in the target form. Since the publication of their seminal study, researchers across a range of second language acquisition subdomains have delved into the factors that regulate the effectiveness of instruction in their respective disciplines. L2 pronunciation is no exception.

L2 pronunciation instruction research can be understood as coming in two waves. The first wave, which is best conceptualized as a precursor to contemporary research, was predominantly concerned with promoting a paradigm shift toward an intelligibility- and comprehensibility-based model. The goal of this model is to help learners develop pronunciation patterns that are easy to understand rather than phonetically nativelike (Levis, 2005, 2020). In an early study on the topic, Munro and Derwing (1995) showed that accented speech is not necessarily difficult to understand, a finding that has been replicated in several follow-up studies (Huensch & Nagle, 2021, 2022; Nagle & Huensch, 2020). The outcome of this first wave was to show that while some non-targetlike pronunciations certainly affect

understanding, not all do, and those that do may have different weights depending on their functional relevance in the target language (Munro & Derwing, 2006).

The second wave of research, which has come into its prime in recent years, has been squarely focused on understanding if pronunciation instruction is effective and, if so, what makes it most effective. There is no question that pronunciation instruction works. Several research syntheses have shown as much (Lee et al., 2015; Saito & Plonsky, 2019; Thomson & Derwing, 2015). However, the question of how instruction can be made most effective—that is, how it can be optimized—remains open. One means of beginning to answer this question is through research synthesis and meta-analysis, which can shed light on the factors that shape the efficacy of instruction through moderator analyses. In a meta-analysis of perception training studies, Sakai and Moorman (2018) reported that perception training leads to small but reliable gains in production. They also found that the proficiency of the learners, the training context, and the total length of the training, among other factors, were significantly related to the amount of production gains at posttest. In another meta-analysis of high variability pronunciation training (HVPT) studies, Zhang et al. (2021b) explored single vs. multi-talker training conditions to determine if multi-talker training is always more beneficial than a single-talker approach. They found an advantage for multi-talker conditions but also reported substantial heterogeneity in the strength of that effect. And in his overview and research synthesis on HVPT, Thomson (2018) outlined several unknowns about how the efficacy of this particular technique can be maximized. These studies clearly demonstrate that optimizing instruction is a central topic in L2 pronunciation research.

Meta-analysis is not the only means of validating findings and assessing the strength of existing evidence. Indeed, meta-analysis works by taking what already exists and uncovering patterns in the literature. Thus, in a meta-analytic approach, studies that are conceptually related but, in many ways, methodologically diverse are collected, coded, and evaluated to discover trends. However, obtaining summary estimates of an effect based on pooled findings may not clarify the precision of those outcomes given the inherent heterogeneity of the studies included in the report pool (Rothstein et al., 2005). Furthermore, meta-analysis cannot by its very nature address gaps in the literature. Simply put, it is not possible to synthesize literature that does not exist in the first place. As a result, although meta-analysis is certainly useful, it must be paired with other forms of data generation that allow studies to be directly compared to one another.

Replication is ideally suited to this endeavor. Close replications, in which one major variable is deliberately modified (Porte & McManus, 2019), generate data that are maximally comparable to the original study by reducing researcher degrees of freedom in methodology and analysis. Approximate replications, following a close replication in sequence by modifying two variables (Porte & McManus, 2019), are explicitly concerned with constructing a systematic program of research that builds confidence in original findings and ensures that any differences in findings are not owing to methodological differences across studies. When the extent of change between an initial study and a replication is larger, as in a conceptual replication that introduces multiple changes to the initial study design simultaneously (Marsden et al., 2018), this serves as a test of the effects or study features under new conditions or in new contexts that helps to determine boundary conditions and where the limits of generalizability lie. Replication studies can also be included in meta-analyses, making aggregated meta-analytic findings more robust. Finally, and perhaps most urgently, by replicating studies in geographically and socio-demographically diverse samples, researchers can contribute to a more ethical and inclusive body of L2 research. For these reasons, meta-analysis and replication go hand in hand and together can potentially reshape the state of the art by illuminating the instructional conditions that lead to the best outcomes for most, if not all, learners.

Many empirical claims need replication, but the question of which studies warrant replication is a not a simple one to address (Isager et al., 2022). For instance, some scholars propose that "[no] special rules for selecting replication studies are needed or even desirable … [because] idiosyncratic interests and methodological expertise guide[d] the original research questions that people pursue[d]" (Zwaan et al., 2018). A less open-ended approach to selecting studies for replication may be to target those that

https://doi.org/10.1017/S0261444823000083 Published online by Cambridge University Press

are "often cited, a topic of intense scholarly or public interest, a challenge to established theories, but … also have few confirmations, imprecise estimates of effect sizes" (Nosek & Lakens, 2013, p. 59). Justifying the choice to replicate an initial study may also be based on a methodological rationale (Hedges & Schauer, 2019) or the weight of currently available evidence (Hardwicke et al., 2018). Indeed, since citation counts rarely offer a "reliable or sufficient motivation for replication" (Marsden et al., 2018, p. 326), studies meriting replication should instead be those that address substantively significant or "theoretically interesting and currently relevant" questions (Mackey, 2012, p. 27). This includes effects and results that are impactful, novel, innovative, theoretically informative, or useful practically. Conversely, studies for which there is a lack of consensus field-wide may also be prime targets for replication because they present inconsistent or surprising findings that have yet to be convincingly corroborated or falsified (Porte & McManus, 2019). We feel that a combination of the above reasons—which tap into notions of value and uncertainty—are useful to inform study selection as long as the rationale for the selection process is transparent.

For this paper, we have selected three studies for replication that offer different perspectives on how L2 pronunciation instruction can be optimized. We have chosen each of these papers precisely because we believe they significantly advance the field conceptually and therefore have broken new ground and point to areas in which replication studies can have a large impact. Martin and Sippel's (2021) study deals with the best means of providing corrective feedback (CF), asking whether students benefit more from providing CF than from receiving it. Olson and Offerman (2021) and Thomson (2012) both focus on optimizing training. Olson and Offerman examined speech production, comparing several visual feedback training paradigms to discover which worked best, whereas Thomson investigated speech perception, combining HVPT with acoustic enhancement techniques to understand if that combination was more beneficial than HVPT alone. In our detailed description of the initial studies for replication below, we present a substantive case for why we believe each study merits replication.

2. First suggested study for replication: Martin and Sippel (2021)

2.1 Background to the study

Martin and Sippel (2021) based their study on the premise that CF is generally an effective way to draw learners' attention to non-salient L2 features and to provide evidence of mismatches between target-like L2 use and current levels of mastery. Their study investigated the effects of teacher and peer corrective feedback (PCF) on the pronunciation development of L2 learners of German, with a specific focus on whether providing feedback, receiving peer feedback, or receiving teacher feedback on pronunciation is more beneficial for students' comprehensibility in their L2 production. Participants in their multi-site sample ($N = 96$) of post-secondary students were all enrolled in a first-year German language course. Two sections of this sample were assigned to each experimental group (Peer Feedback Providers, Peer Feedback Receivers, Teacher Feedback group), and five sections to the Control group. Both a segmental and a suprasegmental target were chosen for the pronunciation tests and training based on their role in impeding L2 intelligibility and reducing comprehensibility for listeners. The segmental target was the grapheme-phoneme correspondence <z> → /ts/ and the suprasegmental target was word stress in German-English cognates (e.g., Konflikt–conflict).

The study took place over a six-week period. In the first week, a pretest was administered to all groups in the form of two controlled production tasks: a word reading task (24 words, more-controlled) and a sentence reading task (12 sentences, less-controlled), each of which contained multiple instances of the segmental and suprasegmental targets. The Peer Feedback groups additionally received metacognitive instruction (see e.g., Sato, 2022), an intervention designed to enhance learners' understanding of the benefits of peer feedback and practice how to provide it. In week two, all participants except those in the Control group received general pronunciation training on the L2 German segmental and suprasegmental target feature. In weeks three, four, and five, all participants, except those in the Control group, went through repeated cycles of completing the production tasks and then either receiving or providing feedback the following week. The Teacher group received feedback

https://doi.org/10.1017/S0261444823000083 Published online by Cambridge University Press

from a teacher, the Peer Provider group gave feedback to a peer, and the Peer Receiver group received feedback from a peer. In the final week, a posttest was administered to all groups. The same words and sentences were used in the production task for both the pretest and posttest, but the order of the words and sentences was randomized.

The results of this study showed that all groups significantly improved their comprehensibility on both segmental and suprasegmental features of the L2, except for the Control group. For the L2 segmental target, larger effects were obtained on the more-controlled word reading task than on the less-controlled sentence reading task. Between-group comparisons revealed that the Peer Feedback Providers improved more than the Peer Feedback Receivers and that receiving feedback from a teacher was superior to receiving feedback from a peer. These analyses suggested that the Peer Feedback Providers had an indirect edge over the Teacher Feedback group. For the suprasegmental target, too, all feedback groups outperformed the Control group in both the more-controlled and less-controlled production tasks. Peer Providers outperformed one other experimental group (i.e., Peer Receivers) in the more-controlled task, but the only significant differences at the sentence level were found between the experimental groups and the Control group. These findings indicated that pronunciation training is better than no training, that pronunciation training combined with feedback is superior to no feedback, and that the Peer Providers benefited more from providing feedback than the Teacher group benefited from receiving instructor feedback.

2.2 Approaches to replication

There are several reasons why Martin and Sippel's study (2021) merits close replication. First, findings about the effects of CF on pronunciation remain inconclusive. Some reviews suggest that CF in pronunciation instruction can have a small to medium effect size on learners' improvement in their pronunciation skills (Lee et al., 2015; Thomson & Derwing, 2015). An important consideration, however, is whether the effects of feedback differ depending on the source. One of the purported strengths of peer feedback is that learners are more actively engaged in the learning process since they are feedback receivers as well as feedback providers. Both aspects of the initial study continue to generate empirical debate and, thus, direct corroborating evidence is needed. Methodologically, the researchers also identified limitations in the design of the initial study that suggest clear avenues for close replication. For instance, their design investigates whether teacher feedback was more effective than peer feedback and whether providers of feedback made more significant gains in their pronunciation than receivers. However, only some of the feedback groups received metacognitive instruction to enhance their understanding of the benefits of feedback, making it unclear whether the effects detected were in fact owing to the feedback provider and not a confound of the metacognitive training as a catalyst for these effects. Additionally, the researchers provided pronunciation training to all but the Control group, making it difficult to conclude that any improvement in learners' pronunciation stemmed from the type of feedback they received and was not a confound of the pronunciation training provided. Finally, the initial study had somewhat counterintuitive findings: all but the Control group improved over time, but only the feedback providers, who were not required to self-correct, re-record, or otherwise engage with the feedback, improved when compared with the other groups. This is an unexpected outcome and suggests that further examination of the question is warranted and would be welcomed by the field.

Here, we would propose a series of close replications looking at one major variable change in each to permit maximum comparison with the original study. First, a close replication could address limitations with the treatment itself. The number of ITEMS in a treatment is central to the duration and intensity of training. For instance, doubling the number of items per session to 48 split across "condition" (i.e., words and sentences) would make the pronunciation training more robust and the tests more sensitive to detecting effects and changes. Related to this variable are the task conditions, which are an important aspect of validating the results of the initial study in a close replication. The initial study administered a set of randomized stimuli repeatedly, and these items differed on two

https://doi.org/10.1017/S0261444823000083 Published online by Cambridge University Press

dimensions: (a) the phonological target (segmental vs. suprasegmental) and (b) the level of control (words = more-controlled production vs. sentences = less-controlled production). A close replication that uses task conditions requiring production of the phonological targets in novel contexts, especially less controlled contexts, would allow researchers to test whether the treatment effects generalize from controlled to spontaneous task conditions, which is a central concern for current L2 pronunciation research (Saito & Plonsky, 2019).

The second variable that might be changed in a close replication is TIME. While the overall window of observation in the initial study (six weeks) worked well, it would no doubt be pedagogically useful to examine whether there are longer-lasting effects than found in the initial study. Adding a delayed test, a necessary component for nearly all experimental designs, would help uncover the durability and transfer of effects (Sippel & Martin, 2022). It would also be important, in a further close replication, for all groups to receive the pronunciation training and for all experimental groups to receive the metacognitive instruction. This would enable the study to determine (1) whether improvement in learners' pronunciation does indeed result from the type of feedback without the added confound of training and (2) whether the effects detected are in fact owing to the feedback provider and not a confound of the metacognitive training as a catalyst for these effects. To link these two first ideas for close replication (items and time), an approximate replication with a pre-post-delayed design could also include spontaneous production (i.e., non-reading) tasks that are counterbalanced across groups to eliminate the potential for task repetition effects.

One of the most intriguing findings of the initial study was that the peer feedback providers, who were not required to self-correct, re-record, or otherwise engage with and uptake the feedback, improved most when compared with the other groups. The initial study discusses the potential mechanisms through which this benefit is realized, namely the increased attention and depth of processing that providing feedback brings. In a close replication, it would also be beneficial to revisit the metacognitive instruction the feedback groups received. This training consisted of a presentation delivered by the researchers focusing on the benefits of CF followed by a brief whole group discussion of why peer feedback might be beneficial. There is much debate in the learning, cognitive, and intervention sciences about the salience of information-based interventions that present new information in text but cannot ensure that the intended impact matches its actual effect. There are other more impactful ways of providing such training that capitalize on mixed modalities and raise participant awareness more deliberately. Including more overt metacognitive instruction in a close replication to compare the effects of feedback provider with metacognitive training, as in the initial study, would help clarify the reasons for this counterintuitive finding.

A further close replication could change the PARTICIPANTS. The initial study sampled first language (L1) users of English enrolled in college-level L2 German courses. This L1–L2 pairing led to specific pronunciation targets. A close replication could recruit learners from new sites learning different L2s that would provide new samples and entirely different pronunciation targets. Given recent calls to look beyond college-level samples in language learning research (Andringa & Godfroid, 2020), a close replication could also sample learners in other age groups in other types of instructional settings. Owing to the nature of the study design and the high reliance on learner involvement in the treatment itself, new samples are also likely to lead to different patterns of findings and shed light on the question of whether providing peer feedback, receiving peer feedback, or receiving instructor feedback on pronunciation is more beneficial for students' comprehensibility in their L2 production. As with any sample considerations, sampling would be best informed by a power analysis to ensure group comparisons are adequately powered to detect desired effect sizes.

Finally, effects are often contingent on the operationalization of outcomes and SCORING. In the initial study, five non-expert L1 raters were recruited in Germany to rate all tests of learner production on "comprehensibility," a general metric of ease of understanding. Raters were asked to rate each word and sentence on a 9-point scale with reference to the criterion question: "How easy or difficult was it for you to understand this utterance?" It is unclear if raters were using just the phonological production or also knew the target and saw it in writing. A close replication could use tiered judgement

https://doi.org/10.1017/S0261444823000083 Published online by Cambridge University Press

that progresses from an intelligibility judgement (e.g., What did you hear?) to a comprehensibility judgement (e.g., How easy was it for you to understand?) to an accuracy judgement (e.g., using a 9-point accuracy scale, which would align with the 9-point comprehensibility scale included in the original research).

3. Second suggested study for replication: Olson and Offerman (2021)

3.1 Background to the study

Olson and Offerman (2021) aggregated data from three of their previous studies, all of which were about the effect of visual feedback training on English speakers' production of voiceless stop consonants /p, t, k/ in Spanish. In each study, learners recorded themselves and printed out visual representations of their production for in-class use, analyzed the characteristics of Spanish stop consonants using acoustic analysis software and compared their production to a native speaker model in class, and practiced producing words in isolation and in utterances. They then rerecorded themselves a few days later and again after four weeks. The goal of the visual feedback paradigm was for learners to notice the difference between their pronunciation and the L2 target and begin adjusting their production toward the norms of Spanish. All three studies showed that the visual feedback training paradigm was highly effective, insofar as all experimental groups improved significantly at posttest. The goal of aggregating data from the studies was to compare the three visual feedback paradigms to one another to determine which paradigm promoted the greatest gains.

The Short paradigm (Olson, 2019) consisted of a single training session following the format laid out previously: pretest recording at home, in-class analysis and comparison, and re-recording within three days (immediate posttest) and after four weeks (delayed posttest). In this study, there were 25 participants who were split into three groups with each group trained using one place of articulation: eight learners completed the visual feedback paradigm focusing on Spanish /p/, 13 on /t/, and 4 on /k/. Participants recorded 30 words in novel utterances per place of articulation (i.e., 30 for /p/, 30 for /t/, and 30 for /k/) at each recording session. Thus, although the target words were always the same, the utterances in which they appeared were not repeated across recordings. Words were embedded in utterance-medial position. Participants also recorded 30 words in isolation, which were repeated at each session. Phonetic context was controlled by combining the word-initial stops with two following vowels: /o/ and /a/. Thus, in this study, Olson tested learning as a result of training via the immediate posttest and retention via the delayed posttest. He also tested two forms of production, both of which would be considered controlled production (Saito & Plonsky, 2019), albeit to varying degrees, with words in isolation representing the most controlled form of production and words in novel utterances a less controlled form of production. By applying a between-subjects design to place of articulation, with participants assigned to work with one type of word-initial stop, he was also able to examine generalization from one target to another (e.g., if participants trained on /p/ they improved their production of /t/ and /k/).

In the Long Simultaneous paradigm (Olson, 2022), 20 participants completed the visual feedback training three times, once per week over a four-week period. Each time, they worked with voiceless stops at all three places of articulation. In this study, the training was scaffolded, such that participants worked with words in isolation, words in utterances, and words in paragraph-length discourse at the first, second, and third session, respectively. As in Olson (2019), participants recorded words in novel utterances and words in isolation. However, the number of stimuli was different (18 per place of articulation) as was the phonetic context in which the stop appeared (stop + /a/, /e/, and /u/). Furthermore, for the words in utterances, in Olson (2019) the target words were utterance-medial but in Olson (2022) they were utterance-initial. Thus, position in the utterance and, as a result, the prosodic context in which the target forms appeared was different across the two studies.

In the Long Sequential approach (Offerman & Olson, 2016), 17 learners participated in three sessions, but training was blocked by place of articulation, such that in the first session, participants worked with /p/, in the second with /t/, and in the third with /k/. There was one session per week,

https://doi.org/10.1017/S0261444823000083 Published online by Cambridge University Press

but words were trained in carrier sentences; there was no scaffolding from words to utterances to discourse. Participants recorded words in novel utterances and words in carrier phrases, but the stimuli characteristics and repetition of stimuli over time were different compared with the other groups. In this case, participants recorded five words per place of articulation, one per vowel (/a/, /e/, /i/, /o/, /u/), and 30 words in a carrier phrase. Participants recorded the same stimuli at the pretest and delayed posttest, but they recorded a different set of stimuli at the immediate posttest.

Comparing the three experimental groups, Olson and Offerman (2021) found no difference between the Short group and the Long Simultaneous group, but both groups were different from the Long Sequential group, which outperformed the other two. This finding suggests that more is not always better. Adding more sessions (Short vs. Long Simultaneous, 1 vs. 3) did not appear to lead to greater gains unless training was blocked by place of articulation (Short and Long Simultaneous vs. Long Sequential). The authors argued that these effects may have emerged because of the processing load of the simultaneous training, where participants worked with all three places of articulation at once.

3.2 Approaches to replication

Olson and Offerman (2021) merits replication for several reasons. For one, it shifts pronunciation research beyond whether instruction is effective to how it can be made most effective, which represents an important step forward in the field. Furthermore, there is considerable debate regarding the training characteristics that maximize learning for all. In the speech perception training literature, although HVPT has been shown to be effective, it is well known that structured variability is most beneficial for all learners (e.g., Perrachione et al., 2011). Here, structured variability means maintaining a high amount of overall variability but reducing trial-by-trial variability by presenting talkers one at a time in a blocked format. Olson and Offerman's (2021) study provides evidence that the same could be true for production training because the Long Sequential group outperformed both the Long Simultaneous and Short groups. Replicating this finding would provide insight into how production training can be optimized, and it would provide evidence of an important parallel between perception and production training paradigms.

There are several aspects of the study that merit attention for the purpose of replication. First, with respect to PARTICIPANTS, because Olson and Offerman's (2021) data sets involved intact classes, it would be worthwhile to conduct a close replication of the study in a new context using a new sample of classes to gain insight into how stable and reproducible the findings are. In doing so, it would be necessary to establish a minimum sample size per paradigm and per class given that group sizes were different in the original study. There are two possibilities for replicating this research with new participants. One option would be to replicate at a single site, in which case (resources permitting) it would be advantageous to assign two or more intact classes to each approach to understand how the classroom context affects findings while maintaining all other major variables intact. Another more ambitious option would be a multi-site design in which the study would be replicated at several sites using several sets of intact classes. Such a design has the potential to shed light on the stability of findings across contexts and the scalability of the approach (Moranski & Zalbidea, 2022). Second, and along the same lines, it would be worthwhile in another study to align ITEMS across paradigms, such that the number of test items, the context in which the target feature appears, and the structure of items over time—that is, whether test items are unique at each data point and counterbalanced with respect to time—are constant across paradigms. Ensuring parity of test items will provide clearer insight into the effect of experimental condition by minimizing the influence of additional variables that varied between groups, such as the following vowel. Furthermore, there may be item-level effects, such that participants tend to produce certain words more accurately than others, which makes it even more important that test items are the same for each training paradigm.

Third, there are several aspects of the TRAINING TASK that can be shored up in a replication. For one, participants in the Short group worked with only one type of voiceless stop, whereas participants in

https://doi.org/10.1017/S0261444823000083 Published online by Cambridge University Press

the Long Simultaneous and Long Sequential groups were exposed to stops at all places of articulation. This design decision was sensible given Olson's (2019) aim of testing generalization to new places of articulation (e.g., the extent to which participants trained on /p/ improved their production of /t/ and /k/). However, for the sake of determining whether the length of the training, blocking of stimuli, or both affect learning, it would be necessary to isolate those effects, holding additional variables constant. In that case, the Short group should receive training on all places of articulation, making it directly comparable to the Long Simultaneous condition. Likewise, there is an important difference with respect to the complexity of the contexts in which the stops were trained. The Long Simultaneous group worked with words in isolation, in utterances, and in discourse across three sessions, whereas the Long Sequential group worked exclusively with words in isolation and in utterances, which were likely mixed within sessions. Working with stops in a range of contexts and blocking training by context may have attenuated gains for the Long Simultaneous group.

Again, for the purpose of identifying the effect of training length and blocking, the contexts in which training stimuli are embedded should be held constant. For instance, all groups could work with words in isolation and words in utterances at each session. Training items should be identical across conditions, which would mean repeating items for the Long Simultaneous and Long Sequential groups. If the long groups receive different training items at each session, then it may not be length, but rather variability in the items—exposure to a greater range of training items—that is responsible for the gains the long groups achieve relative to the Short group. Certainly, these additional variables are interesting in their own right, but to replicate Olson and Offerman's (2021) study on length and blocking, minimizing variability in other aspects of study design is necessary.

Finally, it would be worthwhile to test production using a more spontaneous task, which could provide insight into the extent to which different forms of visual feedback training facilitate the development of controlled and spontaneous production knowledge (Saito & Plonsky, 2019). One example of a spontaneous task is a timed picture description, where participants describe images containing potential target forms. Using such a task would render this an approximate replication. The challenge with such a task is that it would be impossible to strictly control the context in which the target stop occurs. Despite this limitation, using a more spontaneous production task still seems like a valuable addition to the replication.

4. Third suggested study for replication: Thomson (2012)

4.1 Background to the study

HVPT has been established as an effective means of helping learners improve their ability to correctly perceive challenging L2 sounds. A typical HVPT paradigm involves exposing the learner to speech stimuli spoken by many talkers in many phonetic contexts, with the goal of helping the learner attend to the phonetic dimensions that are the most robust cues to the L2 contrast. At the same time, as learners begin to focus on relevant phonetic dimensions, they need to learn to pay less attention to irrelevant ones. Thus, by introducing substantial variability into the stimuli to which the learner is exposed, HVPT essentially forces learners to direct their attention to phonetic dimensions that are robust across speakers and contexts. HVPT is most commonly implemented using an identification task with right/wrong feedback, where the learner hears a spoken stimulus and must map it to a response option. If the learner responds correctly, the program indicates this and advances to the next trial, and if the learner responds incorrectly, they receive feedback and, in many cases, must click on the correct option before the program moves forward. This continues for dozens or even hundreds of trials in a single session, such that over the course of training, the learner potentially identifies a thousand or more stimuli. Generalization tests are an essential means of determining the efficacy of an HVPT paradigm. Three generalization tests are often employed: testing trained items spoken by new talkers, testing untrained items spoken by trained talkers, and testing untrained items spoken by untrained talkers.

In canonical HVPT, the stimuli are completely natural, which means that they are presented to listeners as they were produced by the talkers. Increasingly, however, HVPT has been combined with

https://doi.org/10.1017/S0261444823000083 Published online by Cambridge University Press

other techniques to maximize perceptual learning (Kondaurova & Francis, 2010; Zhang et al., 2021a). Part of the impetus for this type of research is understanding if multi-talker conditions always outperform single-talker conditions (Zhang et al., 2021b) and how such conditions can be made maximally beneficial for all learners (Perrachione et al., 2011). Like other work in this area, Thomson (2012) did not aim to test canonical HVPT. Instead, he sought to test the benefits of perceptually enhancing the vowel stimuli by lengthening them, and he also included several distinct sets of L2 English vowel training stimuli whose relationship to Mandarin, the participants' L1, was variable. Thus, in his study, he examined whether enhancing the stimuli promoted additional gains beyond the baseline HVPT model and if certain sets of L2 sounds responded better to HVPT than others.

Twenty-six L1 Mandarin speakers participated in the study and were assigned to one of three experimental groups: a Lengthened Vowel group ($n = 11$) that received stimuli whose duration had been doubled, a Deselect Vowel group ($n = 4$) that worked with the English vowels that were most likely to be assimilated to Mandarin vowel categories, and a Select Vowel group ($n = 11$) that worked with the vowels that were the least likely to be assimilated to Mandarin. Participants were trained on ten Canadian English vowels. The training stimuli were produced by 20 Canadian English talkers, ten male and ten female, and the testing stimuli were produced by two talkers, one male and one female. Of these two, the male was also one of the talkers included in the training sets, so only the female talker was a new talker for the purpose of testing generalization. Stimuli were recorded and presented in two phonetic environments, /b/ + vowel and /p/ + vowel. Testing included these phonetic environments as well as several others recorded by the new (i.e., the female) talker: /k/ and /g/ + vowel (i.e., vowels presented in a novel stop consonant context) and /s/ and /z/ + vowel (i.e., vowels presented in a novel fricative context).

Participants took part in eight sessions over three weeks, working with 200 trials per session. Response options were ten nautical flags, which were chosen specifically to avoid orthographic representations. These options were presented during an initial familiarization session where participants worked with stimuli produced by a single talker to learn sound-symbol associations. Identification was used for training. On each trial, the stimulus was played, and participants had to click on the correct flag. If they answered correctly, they heard a chirp and the program advanced. If they answered incorrectly, they heard a beep, the correct response flashed, and the trial repeated so that they could select the correct option. Participants completed testing blocks after four and eight sessions (i.e., halfway through and immediately after training) and one month later.

Thomson analyzed both the training and testing data. Regarding training, results showed all vowels improved over time for all groups (except /ʊ/ for the Select Vowel group), but the rate and shape of change varied by vowel and training group, at least descriptively. With respect to testing, for most vowels, participants performed equally well on the trained (i.e., male) and untrained (i.e., female) talkers, which suggests that they were able to generalize to at least one new voice. In terms of generalization to new phonetic contexts, which were only produced by the new talker, participants' vowel identification improved for the trained bilabial stop context (/p/ and /b/) and the untrained velar stop context (/k/ and /g/) but not in the untrained fricative context (/s/ and /z/). Thus, participants had some success at generalizing to an untrained context that was like the trained context but not to an untrained context that was quite different from the one that was trained. Finally, scores at the immediate and delayed posttests were highly correlated with one another, which suggests that performance was mostly maintained over the month-long interval between training and delayed posttesting. It is important to bear in mind that these general patterns mask considerable variability when considering individual vowel targets.

4.2 Approaches to replication

Thomson's (2012) study deals with what has arguably been the most important training paradigm in the speech perception literature. Yet, there are several compelling reasons for why Thomson (2012) should be replicated. First, although studies have generally shown a facilitative effect for high

https://doi.org/10.1017/S0261444823000083 Published online by Cambridge University Press

variability, multi-talker training paradigms, there is substantial heterogeneity in the extent to which high variability conditions outperform their low-variability, single-talker counterparts (Zhang et al., 2021b). There is therefore a basic need to increase the amount of empirical evidence on HVPT and related training paradigms, which "would provide more convincing arguments with adequate statistical power" (Zhang et al., 2021b, p. 4815). Second, beyond solidifying current knowledge on the role of multi-talker training paradigms, much remains unknown about individual differences in the rate and shape of learning during HVPT and the amount of training required to reach ceiling (Thomson, 2018). As a result, we can envision two types of replications in this area: close replications to expand the base HVPT literature and approximate replications to refine our understanding of short- and long-term HVPT learning and its moderators.

In a close replication of Thomson (2012), it would be profitable to increase the number of PARTICIPANTS, such that the participant size per group is equal (e.g., 45 participants with 15 assigned to each group). However, rather than proposing a series of close replications that progressively modify one variable at a time from this study, as we have for the previous papers we selected, we believe replications that introduce more than one change would be appropriate given the fairly advanced state of the art in current HVPT research.

In an approximate replication that introduces two changes to the initial study design simultaneously, we envision two potential paths: (1) increasing the number of SESSIONS and/or the number of TRIALS PER SESSION while maintaining the ten training targets included in the original study, or (2) reducing the NUMBER OF TARGETS while maintaining eight training sessions and 200 trials per session, in line with the original design. On a broad level, these replications could speak to important yet complex questions that have now come to the forefront of the HVPT literature, such as the shape of the learning curve under a variety of training conditions, while holding constant as many of the elements of the initial study as possible.

Both paths to replication would provide meaningful information on how the number of targets and the training structure (i.e., the number of sessions and number of trials per session) interact to regulate perceptual learning over time, which would complement meta-analytic findings on how the number of training targets affects gains at posttest (e.g., Sakai & Moorman, 2018). Furthermore, by manipulating only these characteristics, the replications would remain substantively linked to the original study, allowing for comparisons with the original data. In the spirit of Thomson (2018), these approximate replications can also ask how the number of training targets might regulate the rate and shape of learning over time. Is this development quadratic (i.e., exhibiting curves), cubic (i.e., with rising and falling trajectories), or even nonlinear (i.e., asymptotic, demonstrating ceiling effects)? If HVPT is about variability, then it is easy to imagine that introducing many targets would increase the overall variability of the training paradigm, which could have important repercussions for development. For example, training several targets simultaneously could induce cognitive overload, leading to slower gains over time, especially in an interleaved training format where targets are intermixed across trials. Such an effect might follow from what is known about blocked vs. interleaved multi-talker conditions. That is, while multi-talker conditions are beneficial overall, the best multi-talker format appears to be a blocked one, where talkers are presented one-by-one, such that overall variability is high but trial-to-trial variability is low (Perrachione et al., 2011).

What might such approximate replications look like? For Option 1, where the number of sessions and/or trials per session are increased, it seems reasonable to double either parameter: 16 sessions over six weeks, in which case it would be important to keep the spacing of sessions identical to the original study, or 400 trials per session for eight sessions over three weeks, again maintaining the same inter-session spacing as the original study. For Option 2, it would be sensible to train the most challenging vowels. For most L2 English speakers, those tend to be /i/-/ɪ/ and /æ/-/ɛ/ or /æ/-/ʌ/. Thus, the training set could be halved to five vowels. In either case, we do not believe it would be necessary to replicate Thomson's (2012) Lengthened, Select Vowel, and Deselect Vowel groups because no significant differences emerged in the original study. Not including these training sets would essentially be the second variable change in this approximate replication. Of course, this does not mean that no

https://doi.org/10.1017/S0261444823000083 Published online by Cambridge University Press

differences would emerge in a close replication with a larger sample size (see above), but given the current thrust of HVPT research, focusing on the number of sessions, number of trials per session, or number of training targets seems like a more fruitful approach. It also bears mentioning that these examples are not the only potentially meaningful manipulations of each parameter but rather should be taken as examples that illustrate how researchers can take a principled approach to validating, building upon, and extending the general claims arising out of Thomson's (2012) findings through a systematic series of replications examining the effects of changes in two variables at a time.

5. Conclusion and future directions

Lee et al.'s (2015) meta-analysis revealed a medium effect of L2 pronunciation instruction, as did Sakai and Moorman's (2018) meta-analysis for both learners' L2 perception and L2 production abilities. But what is the nature of this instruction, and what does this training represent in the aggregate? Making sense of such consensus statements (i.e., pronunciation instruction is effective) requires knowing "what pronunciation instruction intends to teach, how its effectiveness is assessed, and how the resulting outcomes are interpreted" (Saito & Plonsky, 2019, p. 4). Given L2 pronunciation scholars' limited resources, it is important to determine at what point the field knows that something works and is ready to move on to new areas. Novelty in designs and results remains a key metric for progress and publication in the field. On their own, however, studies that build on each other in new directions may not be the best approach to uncovering the reality of complex, contingent, and situated effects. While novel approaches are certainly welcome, a research program driven primarily by novelty may underestimate the limits of its designs and analytical methods, may not provide informative boundaries on the applicability of findings, and may not sufficiently address the criteria for deciding when something works and compared to what (Morrison, 2022).

Traditionally, it has been common for researchers to justify their work by observing that "little is known" about a particular subject or relationship. However, in an established field such as ours, this is arguably one of the least persuasive rationales for conducting a study (Porte, 2013). For one, it may be the case that "little is known" because the questions that drive the research itself do not merit scholarly attention. To be clear, there are certainly new areas of study in L2 pronunciation research that await our attention, but the pursuit of innovation can result in a haphazard "data accumulation from an uncoordinated research agenda that complicates the interpretation of results across similar studies and comes at the expense of knowledge construction" (Porte & Richards, 2012, p. 285). Indeed, the result of this approach may beget the death-by-a-thousand-cuts demise of the gradual accumulation of knowledge because the studies upon which such knowledge is built are actually not as comparable as they seem. Replication research is, therefore, an essential part of our field (Porte & McManus, 2019) and can help us accomplish many goals (Marsden et al., 2018). At the very least, carefully revisiting previous findings illuminates blind spots and adds detail to existing knowledge providing a more comprehensive understanding of how effects operate, under which conditions they are maximally effective, and for whom they have the strongest impact.

Part of the enthusiasm for ongoing L2 pronunciation research stems from its promise to provide new insights for practice. In domains of practice (e.g., language instruction), our position is that, at least for the moment, the field should stop innovating and start replicating (or at least, decelerate the pace of innovation and accelerate the pace of replication). Combating the incentive structure that favors innovation over verification requires a parallel model of research (Porte & McManus, 2019). For L2 pronunciation research, this parallel strand of research would revisit initial, stand-alone studies of programs, practices, and policies in order to verify prior findings; contribute complementary data on the reliability, validity, and generalizability of those findings; provide robust methodological checks on analytical practices and research designs; and, thereby, yield coherent and more conclusive evidence about what practices work for which people in which places, and through what processes. Now is an especially opportune time to engage in serious replication work in L2 pronunciation because research in this area has exploded in the last ten years, yielding a rich body of findings

https://doi.org/10.1017/S0261444823000083 Published online by Cambridge University Press

awaiting corroboration. At the same time, we are fully cognizant of the fact that systemic issues, including the current ways in which research is incentivized and rewarded, continue to prioritize innovation over replication. It is therefore incumbent upon all of us, especially senior researchers in the field, to pave the way for replication studies to receive the attention and merit they deserve.

Replication as a research endeavor is not without its challenges, including decisions about what and how to replicate, and whether to replicate recent studies or works that have been influential over time—even if these do not reflect current designs and analytical practices (Isager et al., 2022; Marsden & Morgan-Short, 2023). In educational settings, the primary goal of replication is to increase the interpretability of a study's "tantalizingly incomplete results" and fill in the gaps left by the inevitable limitations of the initial study's methods (Porte & McManus, 2019, pp. 8–9). Substantively, any given study represents an individual data point that provides imperfect and incomplete knowledge. For example, in certain cases, it may still be unclear how pronunciation interventions work or by what mechanisms and processes a pronunciation treatment is beneficial. Many contextual parameters and individual factors add nuance to understanding how the timing, intensity, and modality of an intervention interface with certain learners' cognitive and non-cognitive capacities, their prior learning experience, and existing extramural opportunities for language contact and use. These are promising focal points that replication research can attend to in modified repetitions of the original studies with potential for useful contributions to the field. We believe that by working at these interfaces, replication research can move the field forward by providing clarification and confirmatory evidence while also attending to the scope of pedagogical applications.

Conflict of interest. The authors declare none.

References

Andringa, S., & Godfroid, A. (2020). Sampling bias and the problem of generalizability in applied linguistics. *Annual Review of Applied Linguistics*, *40*, 134–142. doi:10.1017/s0267190520000033

Hardwicke, T. E., Tessler, M. H., Peloquin, B. N., & Frank, M. C. (2018). A Bayesian decision-making framework for replication. *Behavioral and Brain Sciences*, *41*, e132. doi:10.1017/S0140525X18000675

Hedges, L. V., & Schauer, J. M. (2019). More than one replication study is needed for unambiguous tests of replication. *Journal of Educational and Behavioral Statistics*, *44*(5), 543–570. doi:10.3102/1076998619852953

Huensch, A., & Nagle, C. (2021). The effect of speaker proficiency on intelligibility, comprehensibility, and accentedness in L2 Spanish: A conceptual replication and extension of Munro and Derwing (1995a). *Language Learning*, *71*(3), 626–668. doi:10.1111/lang.12451

Huensch, A., & Nagle, C. (2022). Revisiting the moderating effect of speaker proficiency on the relationships among intelligibility, comprehensibility, and accentedness in L2 Spanish. *Studies in Second Language Acquisition*. Advance online publication. doi:10.1017/S0272263122000213

Isager, P. M., van Aert, R., Bahník, Š., Brandt, M., DeSoto, K. A., Giner-Sorolla, R., Krueger, J., Perugini, M., Ropovik, I., van 't Veer, A., Vranka, M., & Lakens, D. (2022). Deciding what to replicate: A decision model for replication study selection under resource and knowledge constraints. *MetaArxiv*. doi:10.1037/met0000438

Kondaurova, M. V., & Francis, A. L. (2010). The role of selective attention in the acquisition of English tense and lax vowels by native Spanish listeners: Comparison of three training methods. *Journal of Phonetics*, *38*(4), 569–587. doi:10.1016/j.wocn.2010.08.003

Lee, J., Jang, J., & Plonsky, L. (2015). The effectiveness of second language pronunciation instruction: A meta-analysis. *Applied Linguistics*, *36*(3), 345–366. doi:10.1093/applin/amu040

Levis, J. (2020). Revisiting the intelligibility and nativeness principles. *Journal of Second Language Pronunciation*, *6*(3), 310–328. doi:10.1075/jslp.20050.lev

Levis, J. M. (2005). Changing contexts and shifting paradigms in pronunciation teaching. *TESOL Quarterly*, *39*(3), 369–377. doi:10.2307/3588485

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

Marsden, E., & Morgan-Short, K. (2023). (Why) are open research practices the future for the study of language learning? *Language Learning*.

Marsden, E., Morgan-Short, K., Thompson, S., & Abugaber, D. (2018). Replication in second language research: Narrative and systematic reviews and recommendations for the field. *Language Learning*, *68*(2), 321–391. doi:10.1111/lang.12286

Martin, I. A., & Sippel, L. (2021). Is giving better than receiving? *Journal of Second Language Pronunciation*, *7*(1), 62–88. doi:10.1075/jslp.20001.mar

Moranski, K., & Zalbidea, J. (2022). Context and generalizability in multisite L2 classroom research: The impact of deductive versus guided inductive instruction. *Language Learning*, *72*(S1), 41–82. doi:10.1111/lang.12487

Morrison, K. (2022). *Taming randomized controlled trials in education: Exploring key claims, issues and debates*. Routledge.

Munro, M. J., & Derwing, T. M. (1995). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning*, *45*(1), 73–97. doi:10.1111/j.1467-1770.1995.tb00963.x

Munro, M. J., & Derwing, T. M. (2006). The functional load principle in ESL pronunciation instruction: An exploratory study. *System*, *34*(4), 520–531. doi:10.1016/j.system.2006.09.004

Nagle, C. L., & Huensch, A. (2020). Expanding the scope of L2 intelligibility research. *Journal of Second Language Pronunciation*, *6*(3), 329–351. doi:10.1075/jslp.20009.nag

Norris, J. M., & Ortega, L. (2000). Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis. *Language Learning*, *50*(3), 417–528. doi:10.1111/0023-8333.00136

Nosek, B., & Lakens, D. (2013). Call for proposals special issue of social psychology on "Replications of important results in social psychology". *Social Psychology*, *44*(1), 59–60. doi:10.1027/1864-9335/a000143

Offerman, H. M., & Olson, D. J. (2016). Visual feedback and second language segmental production: The generalizability of pronunciation gains. *System*, *59*, 45–60. doi:10.1016/j.system.2016.03.003

Olson, D. J. (2019). Feature acquisition in second language phonetic development: Evidence from phonetic training. *Language Learning*, *69*(2), 366–404. doi:10.1111/lang.12336

Olson, D. J. (2022). Phonetic feature size in second language acquisition: Examining VOT in voiceless and voiced stops. *Second Language Research*, *38*(4), 913–940. doi:10.1177/02676583211008951

Olson, D. J., & Offerman, H. M. (2021). Maximizing the effect of visual feedback for pronunciation instruction. *Journal of Second Language Pronunciation*, *7*(1), 89–115. doi:10.1075/jslp.20005.ols

Perrachione, T. K., Lee, J., Ha, L. Y., & Wong, P. C. (2011). Learning a novel phonological contrast depends on interactions between individual differences and training paradigm design. *Journal of the Acoustical Society of the America*, *130*(1), 461–472. doi:10.1121/1.3593366

Porte, G. K. (2013). Who needs replication? *CALICO Journal*, *30*(1), 10–15. doi:10.11139/cj.30.1.10-15

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

Porte, G. K., & Richards, K. (2012). Replication in second language writing research. *Journal of Second Language Writing*, *21*(3), 284–293. doi:10.1016/j.jslw.2012.05.002

Rothstein, H., Sutton, A. J., & Borenstein, M. (2005). *Publication bias in meta-analysis: Prevention, assessment and adjustments*. Wiley.

Saito, K., & Plonsky, L. (2019). Effects of second language pronunciation teaching revisited: A proposed measurement framework and meta-analysis. *Language Learning*, *69*(3), 652–708. doi:10.1111/lang.12345

Sakai, M., & Moorman, C. (2018). Can perception training improve the production of second language phonemes? A meta-analytic review of 25 years of perception training research. *Applied Psycholinguistics*, *39*(1), 187–224. doi:10.1017/s0142716417000418

Sato, M. (2022). Metacognition. In S. Li, P. Hiver, & M. Papi (Eds.), *The Routledge handbook of second language acquisition and individual differences* (pp. 95–109). Routledge.

Sippel, L., & Martin, I. (2022, June 16–18). The effects of peer and teacher feedback on long-term maintenance of gains in L2 pronunciation [Paper presentation]. Pronunciation in Second Language Learning and Teaching Conference, St. Catharines, Ontario, Canada.

Thomson, R. I. (2012). Improving L2 listeners' perception of English vowels: A computer-mediated approach. *Language Learning*, *62*(4), 1231–1258. doi:10.1111/j.1467-9922.2012.00724.x

Thomson, R. I. (2018). High variability [pronunciation] training (HVPT): A proven technique about which every language teacher and learner ought to know. *Journal of Second Language Pronunciation*, *4*(2), 208–231. doi:10.1075/jslp.17038.tho

Thomson, R. I., & Derwing, T. M. (2015). The effectiveness of L2 pronunciation instruction: A narrative review. *Applied Linguistics*, *36*(3), 326–344. doi:10.1093/applin/amu076

Zhang, X., Cheng, B., Qin, D., & Zhang, Y. (2021a). Is talker variability a critical component of effective phonetic training for nonnative speech? *Journal of Phonetics*, *87*, Article 101071. doi:10.1016/j.wocn.2021.101071

Zhang, X., Cheng, B., & Zhang, Y. (2021b). The role of talker variability in nonnative phonetic learning: A systematic review and meta-analysis. *Journal of Speech, Language, and Hearing Research*, *64*(12), 4802–4825. doi:10.1044/2021_JSLHR-21-00181

Zwaan, R. A., Etz, A., Lucas, R. E., & Donnellan, M. B. (2018). Making replication mainstream. *Behavioral and Brain Sciences*, *41*, Article e120. doi:10.1017/S0140525X17001972

Charlie Nagle is Associate Professor in the Department of Spanish and Portuguese at The University of Texas at Austin. He studies second language pronunciation and speech learning, and the linguistic and learner variables that regulate development. He has also published on speaking research methods, including advanced quantitative techniques such as multi-level modeling and longitudinal structural equation modeling. His work has been supported by the National Science Foundation and the Fulbright Commission.

Phil Hiver is Associate Professor in the School of Teacher Education at Florida State University. His research investigates the interface of individual differences and language development with classroom pedagogy. He also writes on innovation in applied linguistics research with a particular focus on open science and methods for studying complex dynamic systems. He is co-editor of the *Routledge handbook of second language acquisition and individual differences* (2022, with S. Li and M. Papi).

Cite this article: Nagle, C., & Hiver, P. (2024). Optimizing second language pronunciation instruction: Replications of Martin and Sippel (2021), Olson and Offerman (2021), and Thomson (2012). *Language Teaching*, *57*(3), 419–432. https://doi.org/10.1017/S0261444823000083