ReCALL (2023), 35: 2, 225–238
doi:10.1017/S0958344023000071

RESEARCH ARTICLE

Beyond replication: An exact replication study of Łodzikowski (2021)

Dennis Foung
The University of British Columbia, Canada (dennis.foung@gmail.com)

Lucas Kohnke
The Education University of Hong Kong, Hong Kong (lucaskohnke@gmail.com)

Abstract

Replication studies have become an emerging line of research in recent decades, including in computer-assisted language learning (CALL). Exact replication, which closely follows a study's protocol, is rare as it is hard to recreate results without establishing a highly controlled environment. However, using data available online, we were able to conduct an exact replication of Łodzikowski's (2021) study, which reported on the use of an allophonic transcription tool by 55 Polish learners of English. Allophonic features are used by native speakers to produce acoustic variants of the same phoneme. The original study offered learners an allophonic transcription tool, examined how they used it and considered its association with phonological awareness. This study extended the original research by addressing the limitations of its regression and transcription analyses. Our findings allowed us to offer several suggestions on (1) how an allophonic transcription tool can be better designed to help learners, (2) how CALL researchers can acquire more data for more useful research and (3) why more replication studies are needed in CALL.

Keywords: allophone; computer-assisted language learning; exact replication; pronunciation; English as a second language; quantitative analysis

1. Introduction

Historically, teaching pronunciation has come in and gone out of fashion, at times marginalised and at other times the centre of language teaching (Pennington & Rogerson-Revell, 2019). While English as a second language (ESL) learners consider pronunciation extremely valuable (Derwing & Rossiter, 2002), instructors often lack the training, resources and confidence to teach it (Bai & Yuan, 2019; Pennington & Rogerson-Revell, 2019), and it tends to play an insignificant role in the curriculum (Gilbert, 2012). ESL learners often have a restricted conception of pronunciation (i.e. striving not for intelligibility, but to pronounce English like a native speaker; Levis, 2005; Scales, Wennerstrom, Richard & Wu, 2006; Thomson & Derwing, 2015). This is a difficult, if not impossible, task because their pronunciation is impacted by the phonological and phonetic features of their first languages (Saito, 2021). In addition, ESL learners often communicate among themselves, so non-native pronunciation is "normal" (Pennycook, 2017). Arguably, therefore, pronunciation instruction should emphasise more realistic goals, such as communicative success (e.g. Derwing, Munro, Foote, Waugh & Fleming, 2014; Isaacs, Trofimovich & Foote, 2018).

Cite this article: Foung, D. & Kohnke, L. (2023). Beyond replication: An exact replication study of Łodzikowski (2021). *ReCALL* 35(2): 225–238. https://doi.org/10.1017/S0958344023000071

Some scholars have proposed that ESL learners should receive explicit pronunciation instruction (Saito & Plonsky, 2019), such as articulatory and/or auditory information about L2 segments ("minimal units of sound defined in phonetic terms"; Pennington & Richards, 1986: 208) and suprasegmental features ("speech sounds larger than phonemes"; Parker & Graham, 2009: 218), as this can improve their pronunciation (e.g. Derwing, et al., 2014; Gao & Weinberger, 2018; Kissling, 2013; Munro & Derwing, 1995; Rogerson-Revell, 2012, 2014; Trofimovich, Kennedy & Blanchet, 2017). Studies have found that segmental phonemes may be easier to learn and teach (Levis, 2005; Saito, 2014), although suprasegmental speech attributes play a significant role in comprehensibility and intelligibility (e.g. Hahn, 2004; Isaacs & Trofimovich, 2012; Kang, Rubin & Pickering, 2010; Saito & Saito, 2017). Lee, Jang and Plonsky's (2015) meta-analysis of 86 studies found that explicit pronunciation instruction had a medium-sized effect. Similarly, Sakai and Moorman's (2018) meta-analysis demonstrated that it led to moderate improvement in the perception and production abilities of learners. However, review articles on the subject are far from conclusive; they often contradict each other and focus on different issues (Zhang & Yuan, 2020).

Generally, phonological awareness is related to the ability to pronounce words in the L2 accurately and is believed to be affected by one's understanding of the phonologies of the L1 and the L2 (Flege & Bohn, 2021) and learner-related factors (Derwing, 2017). The use of the International Phonetic Alphabet (IPA) can help ESL learners understand the English phonological system better and increase their awareness of spelling-to-pronunciation correspondence (Mompean, 2015). Recently, using IPA to teach phonological awareness and sound structure (Wagner & Torgesen, 1987) has gained popularity (Łodzikowski, 2021). This phonemic transcription system represents speech sounds using unique symbols. It helps beginning learners reduce segmental errors that impact intelligibility (e.g. substituting consonantal phonemes; Gao & Weinberger, 2018). A more systematic and narrow form of transcription – allophonic transcription – helps advanced learners improve their pronunciation of salient features, such as fortis plosives or word-initial epenthesis.

The phoneme /t/, which has two allophones, [t] and [t^h], illustrates the difference; these two allophonic variants are features of English. This phoneme is more complex than others because its sound changes depending on its position in the word. The acquisition of allophonic features is essential for L2 learners. Reinisch, Juhl and Llompart (2020) demonstrated that it affects L2 learners' ability to identify regional speakers and varieties. Shea and Curtin (2010) found that L2 learners can acquire allophones based on contextual cues. Evidence from these studies supports the idea that learners need to acquire pronunciation at both the phonetic and allophonic levels.

However, the teaching of allophonic transcriptions continues to be rare in the ESL classroom as (1) teachers lack the time and resources (e.g. transcription tools; Bai & Yuan, 2019; Pennington & Rogerson-Revell, 2019) and (2) learners find that transcription exercises without detailed feedback do not aid in their language acquisition process (Lecumberri & Gallardo, 2003). Several automated transcription tools for teachers are available, such as IPAtranscriptor (Braun, 2020) and Web Transcription Tool (Lecumberri & Gallardo, 2003), which can be used for phonemic transcription activities and selected speech processes (Lecumberri & Gallardo, 2003). Although these tools provide some feedback, the instructor must provide a reference transcription for each activity. Currently, few comprehensive allophonic transcription tools are available for instructors: some have minimal features, while others only cover a few allophonic processes.

The present study aimed to replicate a study by Łodzikowski (2021). This study relied on a custom transcription tool designed for Polish learners of English. Previous studies on Polish learners have adopted a general approach. For example, Rojczyk (2019) examined how Polish learners use allophonic cues to identify word boundaries, finding that they did so with 59% accuracy, comparable to results achieved by Spanish and French learners of English. Other scholars have focused on how Polish learners acquire allophones, especially stops (Rojczyk,

https://doi.org/10.1017/S0958344023000071 Published online by Cambridge University Press

Porzuczek & Bergier, 2013; Schwartz, Balas & Rojczyk, 2014). They consider stops to be particularly important for Polish learners because stop release is mandatory in Polish. Rojczyk *et al.* (2013) found that Polish learners can imitate stops properly, but their ability to do so is affected by distraction. (In their study, the distraction was a flashing digit that appeared on a screen after the participants heard the model pronunciation and before they imitated the stop.) Schwartz *et al.* (2014) found that learners can be affected by their first language in the production of stops, but the mechanism for these patterns needs further research. A more recent study developed machine learning algorithms to assist Polish learners with using stop-related allophonic features and aspiration (Piotrowska *et al.*, 2021). These studies show that certain allophones are particularly important for Polish learners of English, which motivated us to replicate Łodzikowski's (2021) study.

Moreover, as pronunciation is an under-researched area in computer-assisted language learning (CALL), we hope that this replication study will provide insights into the value of allophonic transcription tools and their efficacy in improving the phonological awareness of ESL students. By applying the results of Łodzikowski's (2021) study to CALL, we will provide additional insight and give greater validity to the findings. ESL learners need to understand allophonic features to develop their pronunciation. This study will address how this can be achieved by developing specific, tailor-made tools, which can be applied in different contexts by future researchers.

1.1 Replication studies: A review

Replication studies in applied linguistics can be divided into two groups: internal replication and external replication (Porte & McManus, 2019). External replication adopted the methodology in the original study to produce new data with "new contexts" and "other participants". Internal replication, which is also called "exact replication", is to "re-examine the methods used and conclusions drawn". For example, Nicklin and Plonsky (2020), in an L2 and applied linguistics study, obtained the data set from 104 studies and reanalysed the data with a better outliner detection strategy, and this is one example of internal replication. Luef (2022) obtained a new data set to replicate the research of Siew and Vitevitch (2020) on phonological networks, and that is an example of external replication. Although both external and internal replication studies have achieved some success, conducting an exact replication as an internal replication by another researcher is rarely possible because most research is conducted in a highly controlled setting (Crandall & Sherman, 2016; Porte & McManus, 2019). Past studies may only validate a previously published model with new data (Zeigler, Muzy & Kofman, 2018); however, this process rests on the assumption that the model is a "reliable measure" (Cipriano, Barnes, Kolev, Rivers & Brackett, 2019: 10). Therefore, before validating a model, one must verify it; one must reproduce it as an exact replication using the original data set to confirm that it is reliable. The present exact replication of Łodzikowski's (2021) study is designed to achieve this purpose: to verify the models presented in the original study.

1.2 The original study and suggested approaches to replication

Łodzikowski's (2021) study included two research questions: first, how advanced ESL learners used a supplementary allophonic transcription tool and, second, how their usage of the tool was associated with their declarative phonological awareness. The participants in the original study were 55 Polish English-major undergraduates who used an English transcription tool to supplement a two-semester course on English phonetics and phonology.

The tool allows learners to enter phonetic transcriptions in IPA and see allophonic features. For example, if learners enter /ben/, the tool will show [b̥ẽn]. The tool shows 13 allophonic features by default. These features were based on Cruttenden's (2014) study, which was also used to develop

https://doi.org/10.1017/S0958344023000071 Published online by Cambridge University Press

the course taken by the students in the study. Students accessed the tool through the university learning management system. It has since been posted on a publicly accessible repository (https://bit.ly/phontrans_webapp; Łodzikowski & Aperliński, 2016), and the public can utilise the tool with minimal technical support.

The tool was introduced as a supplementary resource for the two-semester course and the students were not required to use it. However, they were required to complete various transcription activities and assessments, including post-class quizzes, midterm tests and final assessments, all of which required an understanding of allophonic features. The activities and assessments may have incentivised the students to use the tool. A free analytics application, Piwik, was installed to track students' use of the tool (Matomo, 2022). The original study provided sufficient details on how the tracking logs were able to record learners' use of the tool with reasonable accuracy.

Data visualisation was employed to examine the usage data (entries across time and among learners) to answer the study's first research question regarding how the tool was used. Łodzikowski (2021) reported that 91% of the 55 learners used the tool at least once. A total of 3,119 entries were made using the tool over 312 visits. Over the two-semester course, there was a substantial increase in the use of the tool near quizzes and midterm tests and a long period of low use because of the holidays. It is important to note that the raw data for the time-based analysis was not made available, so the current study cannot conduct time-based analysis as the original study did. Besides time-based and learner-based analyses, Łodzikowski conducted brief entry-based analyses. Łodzikowski reported a total of 3,119 entries, with 68 being non-words (i.e. IPA transcriptions entered by learners that were not English words). Among the English word entries, there were 1,105 distinct words and 3,051 total words (a word could be entered multiple times). The 15 most frequently entered words were words on the class transcription worksheets.

To answer the second research question, the original study established three regression models to examine the association between the use of the tool and phonological awareness. Phonological awareness was the target variable for all the regression analyses. This was operationalised using the results of the midterm test (Models 1 and 2) and phonotactics quiz scores (Model 3). A full list of variables and results for three models can be found in Appendices 1a–1c (see supplementary material). The first model included 11 predictors related to learner use and learner background and used the midterm test score as the target variable. It also included the grouping of the students as a random variable. Along with other variables, this mixed-effects model explained 51% of the variance ($R^2 = 0.51$) in phonological awareness, with a student's secondary school writing exam score being a statistically significant predictor ($b = 0.93$, $p < 0.05$). The second model used the same dependent and 11 independent variables but removed grouping as a random variable, making Model 2 a fixed-effects model. Model 2 explained 56% of the variance in phonological awareness, and a student's secondary school writing exam score was once again a statistically significant predictor ($b = 0.93$, $p < 0.05$).

Model 3 was established with the phonotactics quiz score as the target variable and the same list of predictors as Model 2, with the addition of a binary variable (whether learners had entered three or more non-words). Model 3 explained 30% of the variance in phonological awareness, with a student's secondary school oral English score ($b = 3.01$, $p < 0.05$) and mean duration of visit (to the tool) ($b = 0.02$, $p < 0.05$) being statistically significant predictors. Although detailed data processing and cleansing procedures were reported, no details of the assumption-testing procedures were provided in the original study (i.e. the main text), despite indications in the code posted on a separate site. The author concluded that using the tool, as well as visiting it for an additional day, could boost a student's midterm test score. The study also found that entries of non-words could help develop learners' English phonotactics.

https://doi.org/10.1017/S0958344023000071 Published online by Cambridge University Press

Table 1. Similarities and differences of original and current study

Original study: Research objectives	Original study: Data analysis	Current study
Examine how advanced ESL learners used a supplementary allophonic transcription tool	How many learners used the tool?	Same as original study
	When did the learners use the tool?	Not applicable (no data)
	What are the most common transcription entries that learners entered?	Same as original study, except with further investigation of the allophonic features of the entries
Examine potential associations between usage of the tool and levels of declarative phonological awareness among the learners	Mixed-effects regression models: Associations between a midterm test and usage variables, with groups as a random variable	Same as original study, except rerun after removing two variables (see the Results section on why variables were removed)
	Fixed-effects regression models: Associations between the midterm test and usage variables only	Same as original study, except rerun after removing two variables
	Fixed-effects regression models: Associations between a phonotactics quiz score and usage variables	Same as original study, except rerun after removing two variables

2. Methodology

2.1 Research design: Exact replication study

The current study positions itself as an exact replication study that adopted the original data set from Łodzikowski (2021) and followed all of the protocols as closely as possible to recreate the results. It is important to note that although the purpose of an exact replication study is to confirm the results of an original study (Morrison, 2022), we do not argue that the findings of the original study were invalid or unfounded. The purpose of the current study was to extend the original study by discussing some of the limitations that were not emphasised in it and examining some of the findings that could have been further discussed.

Like the original study, the present study examined learner usage data and then developed models to examine the association between the use of the tool and phonological awareness. However, the current study differed from the original study in three ways: (1) we did not recreate the time-based results, (2) we examined the allophonic features of entries, and (3) we revised the models to include fewer variables. See Table 1 for a summary.

2.2 Data retrieval, processing and cleaning

The original study provided two data sets: (1) data on the usage of the transcription tool, organised by the learners; and (2) data about the entries (i.e. transcriptions presented as words) produced by the learners. The data sets were available in the folder "Data" on the website of the tool (https://bit.ly/phontrans_analysis). The previous section has provided more details about the procedures and participants in the original study.

The first data set, recording learners' use of the transcription tool, did not require any processing. The original study reported that detailed screening and cleansing procedures were used on the data concerning the use of the tool by the 55 learners, so no further processing or cleansing was performed. The data set included variables for each learner that were relevant to the current study, including gender, prior achievements (e.g. secondary school writing and oral exam scores), visiting the tool at least once (binary), raw number of visits (i.e. visiting twice on Monday counted as two), number of visits on distinct days (i.e. visiting twice on Monday counted

https://doi.org/10.1017/S0958344023000071 Published online by Cambridge University Press

as one), raw number of inputs (i.e. entering the same word twice counted as two), number of distinct inputs (i.e. entering the same word twice counted as one), mean duration of visit (in seconds), number of visits within seven days before assessment, number of visits within one day before assessment, entering three or more non-words (binary), midterm test score, and phonotactic quiz score.

The second data set had to be processed to examine the allophonic features. We removed the entries with non-words (as no details were provided for them) and those that were only entered once. We then converted the remaining 541 words into IPA using the online version of the Cambridge English Dictionary (https://dictionary.cambridge.org/) to create a transcription (with the UK pronunciation). Next, we entered the transcription into the tool used in the original study to obtain an allophonic transcription. Finally, we labelled each allophonic transcription with the relevant allophonic features. For example, the word "bead" was first transcribed as /bid/. Then, the tool converted that to [b̥id̥] with one feature: devoicing. To ensure that we did not misunderstand the diacritic, we examined the original Java file transcriptor.js as part of the tool. As the second data set was subject to exploratory analysis (i.e. data visualisation), no further processing or cleansing was needed.

2.3 Data analysis

As shown in Table 1, three types of data analysis were conducted in the current study to replicate the results of the original study. To initially examine the usage data by learner, simple descriptive statistics were produced using MS Excel; the transcription entries were processed and analysed. To examine the association between phonological awareness and the use of the transcription tool, regression analyses were conducted.

To ensure the validity of the regression analysis, we first checked whether there were enough samples, according to the criterion suggested by Tabachnick and Fidell (2013). The required sample size was $50 + 8$ * the number of predictors. With only 55 learners, the original study did not have a large enough sample. An insufficient sample size results in the overfitting of the data (i.e. inability to generalise the results to the population). Although we could have decided to end the study here, we believed it was still meaningful to examine the regression models using an exploratory approach.

Next, we conducted three rounds of analysis with the given target variable and predictors (plus the random variable in Model 1). See Appendices 1a–1c for details about these models. As in the original study, we adopted the "enter" approach, including all of the predictors (i.e. not excluding any for non-significance). To further confirm the validity of the analyses, multicollinearity was examined by constructing a correlation matrix with all of the quantitative variables. We checked the sample size and confirmed the presence of multicollinearity while running the analysis to give us a better understanding of the results of the original study. We adopted the criterion suggested by Nimon (2018) – $r > 0.5$ – when examining the correlation matrix. Unfortunately, two of the variables (raw number of visits and number of visits on distinct days) were highly correlated with other variables. Therefore, we decided to remove those two variables and rerun the three models with the same parameters. Details about these decisions can be found in the Results section. We examined the normality of the residuals and found no irregularities in any of the models. To evaluate the models, we examined the R^2, the overall model adequacy, for Models 2 and 3, followed by the significance of the individual predictors.

All of the initial regression analyses were conducted in R (Version 4.0.3; The R Foundation, n.d.). The original study also used R for data analysis and its codes were posted online (https://bit.ly/phontrans_analysis). We believed that the same purposes and results could be achieved with the built-in lm() (for fixed-effects model) command and the lme() command (for mixed-effects model) from the nlme library (Pinheiro *et al.*, 2022). The analyses were conducted using those

Table 2. Summary of allophonic features of entries

Allophonic features	Percentage
Devoicing	45.29%
Pre-fortis clipping	41.03%
Nasalisation	27.54%
Labialisation	19.41%
Velarization	18.30%
Aspiration	14.42%
Vowel retraction before dark L	8.13%
Inaudible plosion	5.18%
Consonant retraction	4.99%
Consonant advancement	3.70%
Nasal plosion	1.66%
Lateral plosion	1.48%
GOAT allophony	0.00%

commands instead of those used in the original study. Also, the glance() command from the broom library ("Introduction to broom", 2022) was used to reproduce the R^2.

3. Results

3.1 Transcription study patterns

As this was an exact replication study, its first aim was to examine how advanced ESL learners used the supplementary allophonic transcription tool. We examined some learner-based usage data and then entry-based data using the data sets provided by the original author. As reported in the original study, 90.90% ($n = 50$) of the learners used the tool at least once. We found that the average number of transcriptions entered per student was 51.78, with the highest being 240 and the lowest 1 (among those who did use the tool). There were 270 visits, with a median time per visit of 16.02 minutes. These results seemed to differ from those of the original study. A summary is presented in Table 2. A further examination of the usage data showed that three learners (in addition to the five who did not use the tool) spent a total of 0 seconds on the tool in their one or two visits. The original paper reported that 16 visits lasted 0 minutes, but we found only four visits that lasted 0 minutes. Generally, these results align with those of the original study despite minor discrepancies, which did not come as a surprise to the authors.

To better understand the usage of the tool, the current study attempted to extend the past study by further analysing the transcriptions that the students entered. This data set was posted, along with the usage data; it contained 3,051 entries comprising 1,100 words and 68 entries of non-words. To extend the original study, we entered the transcriptions into the tool again, excluding non-word entries and words entered only once ($n = 559$), and analysed the allophonic features identified by the tool. The remaining 541 words (2,451 entries) had an average of 1.99 allophonic features each; 7.76% had no allophonic features (e.g. cheap [ʧip] and get [gɛt]). Some had more, such as consequence [ˈk̠ʰɑ̃nsə̥̆kʷwə̃ns], with seven features, and transcriptor [ˈt̠ɹ̥æ̃nˌskɹɪ̆p̚təɹ], with six.

Among the 13 default features (Table 2), the one found in the highest proportion of words was devoicing (45.29%), followed by pre-fortis clipping (41.03%). No entries had GOAT allophony

Table 3. Correlation matrix of related variables

	Raw No. of visits	No. of visits on distinct days	Raw No. of inputs	No. of distinct inputs
Raw No. of visits	1	0.98*	0.74*	0.75*
No. of visits on distinct days		1	0.77*	0.78*
Raw No. of inputs			1	0.99*
No. of distinct inputs				1

*$p < 0.05$.

(from [əʊl] to [ɒʊɫ]) and fewer than 2% had nasal plosion (1.66%) or lateral plosion (1.48%). A summary is provided in Table 2. These findings provide preliminary indications of how learners use the tool and deserve further discussion.

3.2 Association between tool use and phonological awareness level

The second aim of the study was to examine the association between the use of the tool and phonological awareness. As reported in the Data Analysis section, the sample size seemed too small to produce reliable results, but we chose to reproduce the results using the data available. In both the original and present studies, three models were tested; they contained predictors related to tool usage and midterm test scores as the dependent variable (see Appendices 1a–1c for details). The first was a mixed-effects model that included 11 predictors and the grouping of the students as a random variable. The second was a fixed-effects model that used the same 11 predictors as the first model. The third model included one additional predictor, whether the learners had entered three or more non-words, for a total of 12 predictors. A summary of the three models and the strength of their predictors can be found in Appendices 1a–1c.

The first mixed-effects model showed an adequate R^2 that explained 51% of the variance. The second fixed-effects model produced similar results, with an R^2 of 56%. In both models, a student's secondary school writing exam ($B = 0.93$, $p < 0.05$), the number of visits on distinct days ($B = 2.62$, $p < 0.05$) and the number of distinct inputs ($B = 0.16$, $p < 0.05$) were statistically significant predictors of the midterm test score. In the third model, which included non-word entries, a student's secondary school oral exam score ($B = 3.01$, $p < 0.05$), the number of visits on distinct days ($B = 0.47$, $p < 0.05$), number of distinct inputs ($B = 0.16$, $p < 0.05$) and the mean visit duration ($B = -0.52$, $p < 0.05$) were statistically significant predictors. ANOVA was used to examine the overall adequacy of the second and third models. The second model was adequate ($F = 5.03$, $p < 0.05$) but the third was not ($F = 1.05$, $p > 0.05$). It was encouraging to note that our research managed to reproduce the results of the original study. Forty-five of the 46 parameters computed in both studies were identical (the one exception being BIC (Bayesian Information Criterion) for the first model).

Multicollinearity must be examined after establishing the models. As discussed in the Data Analysis section, when multicollinearity is present, statistically significant predictors can appear non-significant due to large standard errors (Tabachnick & Fidell, 2013). In this study, a pairwise correlation of 0.8 or above was considered a concern (Nimon, 2018). The correlation matrix of four related variables can be found in Table 3. These four variables were chosen because they were highly relevant to each other in theory. Among the four variables, two pairs of predictors had a correlation of 0.8 or above (the number of inputs and the number of distinct inputs; the number of visits and the number of visits on distinct days; see Methodology for examples of how these variables differ). This result was expected as the variables are the same in nature. The presence of multicollinearity suggested that the original models might have had limitations.

There are three common ways to address multicollinearity problems: using a larger sample, removing variables or transforming variables. In acknowledgement of the sample-size limitations, we decided to re-establish the regression equations after removing the two highly correlated predictors. As they measured similar constructs (i.e. distinct entries vs. all entries; distinct visits vs. all visits) and the author of the original study acknowledged the importance of counting the distinct entries, we decided to remove the "all visits" and "all entries" predictors. See Appendices 1a–1c for the three recreated regression models.

The re-established regression models were comparable to the original ones in terms of overall adequacy, R^2 and statistically significant predictors. However, the strength of the predictors for input and visit dropped by approximately 30% to 50%, except for the visit count in Model 3, which increased by 19%. Thus, two observations made in the original study need to be revised: After controlling for all other factors, visiting the tool once resulted in a 5.44% increase in midterm test score (originally, 5.03%) and each distinct day of visiting the tool resulted in a 1.73% score increase. As in the original study, most of the predictors were not statistically significant. Only the secondary school exit exams that tested written and spoken English were significant predictors of the midterm test score. The secondary school exit exam of spoken English and the average visit duration were statistically significant predictors of the phonotactics quiz score. Although the results in the recreated regression analyses seemed comparable, the sample-size limitations and the significance of predictors for CALL studies deserve further discussion.

4. Discussion

4.1 The value of exact replication studies

One of the aims of the current study was to conduct a specific type of replication for Łodzikowski's (2021) study, an exact replication study. The results suggested that the reproduced results were highly comparable to those obtained in the original study, with only minor differences. However, through the replication process, we identified certain limitations in the original study – sample size and collinearity – that could have been discussed further. It is important to emphasise once again that we, by no means, are claiming that the arguments in the original study were unfounded. Instead, we would like to extend the original study by discussing some potential limitations. Although the regression models in the current study worked with the data set and the sample size was representative of the population, the sample-size requirements could have been further discussed in the limitations section. In the same vein, collinearity was to be expected for the given variables. In other words, while it is technically possible to recreate the results of the original study using the same data analysis methods, we argue that results in the original study should be read cautiously and as those of an exploratory study. While past replication studies have measured the success of replication by the extent to which the new results were similar to or different from the original results, this study was different: the results were reproducible but not successful. This finding offers a unique perspective on replication studies.

The current study, as a specific type of replication study, an exact replication study, demonstrates the value of conducting exact replication studies. We used the process of identifying and discussing limitations to identify "undiscovered bugs" in the original study, as described by Miłkowski, Hensel and Hohol (2018). Models such as the three in the current study can be "reliable estimate[s]" only if they can be verified (Cipriano *et al.*, 2019: 10). Conducting a direct and conceptual replication that attempts to follow the procedures of the original study and recreate its results (Porte & McManus, 2019) can contribute to the body of knowledge by showing how the results differ from or are similar to the original findings. If a conceptual replication were to be conducted for the current study, more could be discovered about the association between transcription tool use and phonological awareness. In the current study, we contributed to the

https://doi.org/10.1017/S0958344023000071 Published online by Cambridge University Press

body of knowledge in a different but equally meaningful way by verifying the findings of the original study and providing support for the evidence that it contributed to the literature. From a broader perspective, the results of this study echo the need for more exact replication studies in education research (Morrison, 2022). To achieve this goal, more data should be made available for educational studies (Hiver & Al-Hoorie, 2020).

4.2 Phonological awareness: Extending the original study

We extended the original study and provided additional insights into the acquisition of allophones and phonological awareness using the existing data that was shared with us. The original study presented only the 15 most frequently entered words and reported that they came from the in-class transcription worksheets. As mentioned by the author, some entries may have been overrepresented, but the list of words still provided insights into the ways in which Polish learners used a transcription tool to acquire allophones. The learners in this study had some awareness of allophonic features from their classroom instruction, so the transcription entries offered insights into the words or phonemes whose corresponding allophonic features they wanted to know. The students were more interested in some allophonic features than others. It is not surprising to see that plosive release features, including devoicing, aspiration and inaudible release, were quite common in the words learners entered (accounted for 64.88% of entries analysed). Although this could be an artefact of the course materials, another logical explanation is that Polish students want to learn more about the allophonic features of stops. As release is mandatory in Polish (Schwartz *et al.*, 2014), we argue that learners may want to find out more about the contrasting allophonic features of English. The Polish students, as learners of English, may be aware of the differences. Alternatively, their teachers may have drawn their attention to the phonological differences between English and Polish, motivating them to search for more information using the tool. Another interesting observation was that some allophonic features – for example, GOAT allophony – received little attention from learners. Perhaps they had not seen examples of such allophones and thus could not enter valid transcriptions to explore them. For instance, there are few examples of GOAT allophony (e.g. [əʊl]) compared to the number of examples of aspirated plosives (e.g. pat [p^{h}æ̆t]) and unaspirated plosives (e.g. spark [spɑɹk]). To try out the allophonic features, the learners may have typed [əʊl] to see its change to [ɒʊɫ]; however, this entry would have been classified as a non-word in the data set.

To further help learners increase their phonological awareness, especially that of allophonic features, more features could be introduced to similar transcription tools. First, to further address the need to investigate L2-specific allophonic features (e.g. plosive-related features in this study), a transcription tool could incorporate audio so students could hear words pronounced correctly. Rojczyk *et al.* (2013) found that imitation without distraction (e.g. by a tool in front of the computer) helped Polish learners of English acquire allophonic features; providing audio is the logical first step of imitation. Also, to address the problem that there were not enough examples of the less common allophonic features, transcription tools could be designed to allow users to type words instead of IPA symbols. For example, typing the word "swollen" could generate the phonetic transcript, /swəʊlən/, and the corresponding transcription with allophonic features, [s^{w}wɒʊl͠ən]. These suggestions, which are based on the findings of the current study, could help learners increase their phonological awareness and learn the allophonic features of languages.

4.3 Implications for CALL research

The data analysis of the original study appears to have had certain limitations because of the limited sample size. Although the sample was representative of the learners (it included 78.57% of the students in the course that was studied), the number of data points for the

https://doi.org/10.1017/S0958344023000071 Published online by Cambridge University Press

inferential statistical analysis was limiting. An examination of the usage data showed that there were only 270 visits to the tool (versus 320 reported in the original study) from December to mid-June, which means 53 visits per month for 55 learners. As in past studies of CALL tools, some learners may have been more active and used the tool more frequently than other learners (authors). Therefore, the original study examined the behaviours of a very limited number of learners (despite being representative in the context of a blended learning course at a Polish university). As a result, its generalisability to other contexts is limited. From a statistical perspective, the use of regression requires an adequate sample size regardless of how representative a given sample is, so the use of multiple linear regression did not build a strong argument for the association between the use of the CALL tool and phonological awareness. Many CALL studies using multiple linear regression have larger sample sizes. For example, a study by Goh, Sun and Yang (2020) considered six variables with 268 essays and one by Hegelheimer and Tower (2004) considered four variables with 94 students. Therefore, it is not impossible for CALL studies to acquire sufficiently larger samples.

Despite these limitations, we see opportunities for future CALL research. We recommend increasing the number of data points for each learner so that tool usage can be assessed from a range of perspectives. One possible way to do this would be to expand the data points acquired from the tool. The original study reported that learners accessed the tool via a Moodle link and that their usage was tracked with Piwik. Therefore, it would be possible to obtain data on more learner-specific behaviours for analysis. For example, researchers could examine what transcriptions the learners entered, when they entered them, and how they performed on their midterm test on questions that assessed the corresponding allophonic features. This could tell us, for example, if a learner who examines nasalisation frequently before a midterm test (e.g. with multiple distinct inputs or visits) will have better phonological awareness of nasalisation and perform better on related questions on the midterm. A range of exploratory analysis strategies, such as data visualisation, could be used to gather evidence that will be useful for CALL researchers and practitioners.

Another way to collect more data would be to adopt a proactive approach. Given the six-month duration of the course, learners could be invited to complete instruments on phonological awareness before the use of the tool, two months after it was made accessible, before the midterm test and after the phonotactic quiz. With multiple self-reported data points from learners, it would be possible to track changes in learners' phonological awareness over time and look for associations between these changes and their performance on the midterm test. Although such proactive data collection approaches may require complex ethical clearance procedures, they could generate better evidence for CALL researchers and practitioners.

5. Conclusion

This paper reported a unique exact replication study within CALL that attempted to recreate the results of Łodzikowski's (2021) study of a transcription tool. The original study examined the use of an allophonic transcription tool and its association with phonological awareness. In reproducing this study, we extended it by examining its limitations. To a certain extent, we recreated the results of the original study, but we concluded that its models would have been more accurate if there had been fewer variables due to multicollinearity. Also, we believe that the results of the original study should be considered exploratory due to the small sample size. The extension of the study from this perspective offers other CALL researchers the opportunity to consider how they could acquire better data sets using online tools. It also provides additional information on learners' interest in the allophonic process and insights into teaching allophones. Finally, the present study provides new insights into the role of exact replication studies in contrast to other replication studies. Journals and publishers should encourage the sharing of data sets so that more exact replication studies are possible.

https://doi.org/10.1017/S0958344023000071 Published online by Cambridge University Press

Supplementary material. All appendices are available on Figshare: https://doi.org/10.6084/m9.figshare.22109027
To also view supplementary material referred to in this article, please visit https://doi.org/10.1017/S0958344023000071

Data. The data used in this study is made available by the original author. See Methodology for details.

Ethical statement and competing interests. This paper adopted an existing data set being made available online. The authors declare no competing interests.

References

Bai, B. & Yuan, R. (2019) EFL teachers' beliefs and practices about pronunciation teaching. *ELT Journal*, 73(2): 134–143. https://doi.org/10.1093/elt/ccy040

Braun, A. (2020) *IPAtranscriptor*: A Python program for narrow phonetic transcription for blind and sighted linguists. *Journal of the International Phonetic Association*, 50(2): 193–198. https://doi.org/10.1017/S0025100318000233

Cipriano, C., Barnes, T. N., Kolev, L., Rivers, S. & Brackett, M. (2019) Validating the emotion-focused interactions scale for teacher–student interactions. *Learning Environments Research*, 22(1): 1–12. https://doi.org/10.1007/s10984-018-9264-2

Crandall, C. S. & Sherman, J. W. (2016) On the scientific superiority of conceptual replications for scientific progress. *Journal of Experimental Social Psychology*, 66: 93–99. https://doi.org/10.1016/j.jesp.2015.10.002

Cruttenden, A. (2014) *Gimson's pronunciation of English*. Abingdon: Routledge.

Derwing, T. M. (2017) The role of phonological awareness in language learning. In Garrett, P. & Cots, J. M. (eds.), *The Routledge handbook of language awareness*. New York: Routledge, 339–353. https://doi.org/10.4324/9781315676494

Derwing, T. M. & Munro, M. J. (2005) Second language accent and pronunciation teaching: A research-based approach. *TESOL Quarterly*, 39(3): 379–397. https://doi.org/10.2307/3588486

Derwing, T. M., Munro, M. J., Foote, J. A., Waugh, E. & Fleming, J. (2014) Opening the window on comprehensible pronunciation after 19 years: A workplace training study. *Language Learning*, 64(3): 526–548. https://doi.org/10.1111/lang.12053

Derwing, T. M. & Rossiter, M. J. (2002) ESL learners' perceptions of their pronunciation needs and strategies. *System*, 30(2): 155–166. https://doi.org/10.1016/S0346-251X(02)00012-X

Flege, J. E. & Bohn, O.-S. (2021) The revised speech learning model (SLM-r). In Wayland, R. (ed.), *Second language speech learning: Theoretical and empirical progress*. Cambridge: Cambridge University Press, 3–83.

Gao, Z. & Weinberger, S. (2018) Which phonetic features should pronunciation instructions focus on? An evaluation on the accentedness of segmental/syllable errors in L2 speech. *Research in Language*, 16(2): 135–154. https://doi.org/10.2478/rela-2018-0012

Gilbert, J. B. (2012) *Clear speech: Pronunciation and listening comprehension in North American English* (4th ed.). New York: Cambridge University Press.

Goh, T.-T., Sun, H. & Yang, B. (2020) Microfeatures influencing writing quality: The case of Chinese students' SAT essays. *Computer Assisted Language Learning*, 33(4): 455–481. https://doi.org/10.1080/09588221.2019.1572017

Hahn, L. D. (2004) Primary stress and intelligibility: Research to motivate the teaching of suprasegmentals. *TESOL Quarterly*, 38(2): 201–223. https://doi.org/10.2307/3588378

Hegelheimer, V. & Tower, D. (2004) Using CALL in the classroom: Analyzing student interactions in an authentic classroom. *System*, 32(2): 185–205. https://doi.org/10.1016/j.system.2003.11.007

Hiver, P. & Al-Hoorie, A. H. (2020) Reexamining the role of vision in second language motivation: A preregistered conceptual replication of you, Dörnyei, and Csizér (2016). *Language Learning*, 70(1): 48–102. https://doi.org/10.1111/lang.12371

Introduction to broom. (2022) https://cran.r-project.org/web/packages/broom/vignettes/broom.html

Isaacs, T. & Trofimovich, P. (2012) Deconstructing comprehensibility: Identifying the linguistic influences on listeners' L2 comprehensibility ratings. *Studies in Second Language Acquisition*, 34(3): 475–505. https://doi.org/10.1017/S0272263112000150

Isaacs, T., Trofimovich, P. & Foote, J. A. (2018) Developing a user-oriented second language comprehensibility scale for English-medium universities. *Language Testing*, 35(2): 193–216. https://doi.org/10.1177/0265532217703433

Kang, O., Rubin, D. & Pickering, L. (2010) Suprasegmental measures of accentedness and judgments of English language learner proficiency in oral English. *The Modern Language Journal*, 94(4): 554–566. https://doi.org/10.1111/j.1540-4781.2010.01091.x

Kissling, E. M. (2013) Teaching pronunciation: Is explicit phonetics instruction beneficial for FL learners? *The Modern Language Journal*, 97(3): 720–744. https://doi.org/10.1111/j.1540-4781.2013.12029.x

Lecumberri, M. L. & Gallardo, F. (2003) English FL sounds in school learners of different ages. In García Mayo, M. P. & García Lecumberri, M. L. (eds.), *Age and the acquisition of English as a foreign language*. Clevedon: Multilingual Matters, 115–135.

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

Levis, J. M. (2005) Changing contexts and shifting paradigms in pronunciation teaching. *TESOL Quarterly*, 39(3): 367–377. https://doi.org/10.2307/3588485

https://doi.org/10.1017/S0958344023000071 Published online by Cambridge University Press

Łodzikowski, K. (2021) Association between allophonic transcription tool use and phonological awareness level. *Language Learning & Technology*, 25(1): 20–30. https://hdl.handle.net/10125/44748

Łodzikowski, K. & Aperliński, G. (2016, December 1–3) *Usage patterns of an online allophonic transcriptor* [Paper presentation]. 10th International Conference on Native and Non-native Accents of English, Łódź, Poland.

Luef, E. M. (2022) Growth algorithms in the phonological networks of second language learners: A replication of Siew and Vitevitch (2020a). *Journal of Experimental Psychology: General*, 151(12): e22–e44. https://psycnet.apa.org/doi/10.1037/xge0001248

Matomo. (2022) About us. https://matomo.org/about/?footer

Miłkowski, M., Hensel, W. M. & Hohol, M. (2018) Replicability or reproducibility? On the replication crisis in computational neuroscience and sharing only relevant detail. *Journal of Computational Neuroscience*, 45(3): 163–172. https://doi.org/10.1007/s10827-018-0702-z

Mompean, J. A. (2015) Phonetic notation in foreign language teaching and learning: Potential advantages and learners' views. *Research in Language*, 13(3): 292–314. https://doi.org/10.1515/rela-2015-0026

Morrison, K. (2022) *Replication research in education: A guide to designing, conducting, and analysing studies*. London: Routledge. https://doi.org/10.4324/9781003204237

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. https://doi.org/10.1111/j.1467-1770.1995.tb00963.x

Nicklin, C. & Plonsky, L. (2020) Outliers in L2 research in applied linguistics: A synthesis and data re-analysis. *Annual Review of Applied Linguistics*, 40: 26–55. https://doi.org/10.1017/S0267190520000057

Nimon, K. (2018) Multicollinearity. In Frey, B. B. (ed.), *The SAGE encyclopedia of educational research, measurement, and evaluation*. Thousand Oaks: SAGE Publications. https://doi.org/10.4135/9781506326139

Parker, R. & Graham, T. (2009) *The phonology of English: An introduction for teachers of ESOL* (Rev. ed.). Brighton: ELB Publishing.

Pennington, M. C. & Richards, J. C. (1986) Pronunciation revisited. *TESOL Quarterly*, 20(2): 207–225. https://doi.org/10.2307/3586541

Pennington, M. C. & Rogerson-Revell, P. (2019) *English pronunciation teaching and research: Contemporary perspectives*. Basingstoke: Palgrave Macmillan.

Pennycook, A. (2017) The cultural politics of English as an international language. London: Routledge. https://doi.org/10.4324/9781315225593

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., EISPACK authors, Heisterkamp, S., Van Willigen, B., Ranke, J. & R Core Team. (2022) *Package 'nlme'*. https://cran.r-project.org/web/packages/nlme/nlme.pdf

Piotrowska, M., Czyżewski, A., Ciszewski, T., Korvel, G., Kurowski, A. & Kostek, B. (2021) Evaluation of aspiration problems in L2 English pronunciation employing machine learning. *The Journal of the Acoustical Society of America*, 150(1): 120–132. https://doi.org/10.1121/10.0005480

Porte, G. & McManus, K. (2019) What kind of replication should you do? From the inside, looking out: Initial critique and internal replication. In Porte, G. & McManus, K. (eds.), *Doing replication research in applied linguistics*. New York: Routledge, 48–68. https://doi.org/10.4324/9781315621395

Reinisch, E., Juhl, K. I. & Llompart, M. (2020) The impact of free allophonic variation on the perception of second language phonological categories. *Frontiers in Communication*, 5: 1–14. https://doi.org/10.3389/fcomm.2020.00047

The R Foundation. (n.d.) *The R project for statistical computing*. https://www.r-project.org/

Rogerson-Revell, P. (2012) Can or should we teach intonation? *Speak Out! IATEFL Pronunciation SIG Newsletter*, 47: 16–20. https://pronsig.iatefl.org/journal/

Rogerson-Revell, P. (2014) Pronunciation matters: Using English for international business communication. In van den Doel, R. & Rupp, L. (eds.), *Pronunciation matters: Accents of English in the Netherlands and elsewhere*. Amsterdam: VU University Press, 137–159.

Rojczyk, A. (2019) Nonnative perception of allophonic cues to word boundaries: *Lou spills* versus *loose pills* for speakers of Polish. *Language Acquisition*, 26(1): 97–105. https://doi.org/10.1080/10489223.2018.1433672

Rojczyk, A., Porzuczek, A. & Bergier, M. (2013) Immediate and distracted imitation in second-language speech: Unreleased plosives in English. *Research in Language*, 11(1): 3–18. https://doi.org/10.2478/v10015-012-0007-7

Saito, K. (2014) Experienced teachers' perspectives on priorities for improved intelligible pronunciation: The case of Japanese learners of English. *International Journal of Applied Linguistics*, 24(2): 250–227. https://doi.org/10.1111/ijal.12026

Saito, K. (2021) What characterizes comprehensible and native-like pronunciation among English-as-a-second-language speakers? Meta-analysis of phonological, rater, and instructional factors. *TESOL Quarterly*, 55(3): 866–900. https://doi.org/10.1002/tesq.3027

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. https://doi.org/10.1111/lang.12345

Saito, Y. & Saito, K. (2017) Differential effects of instruction on the development of second language comprehensibility, word stress, rhythm, and intonation: The case of inexperienced Japanese EFL learners. *Language Teaching Research*, 21(5): 589–608. https://doi.org/10.1177/1362168816643111

https://doi.org/10.1017/S0958344023000071 Published online by Cambridge University Press

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. https://doi.org/10.1017/S0142716417000418

Scales, J., Wennerstrom, A., Richard, D. & Wu, S. H. (2006) Language learners' perceptions of accent. *TESOL Quarterly*, 40(4): 715–738. https://doi.org/10.2307/40264305

Schwartz, G., Balas, A. & Rojczyk, A. (2014) Stop release in Polish English — Implications for prosodic constituency. *Research in Language*, 12(2): 131–144. https://doi.org/10.2478/rela-2014-0006

Shea, C. E. & Curtin, S. (2010) Discovering the relationship between context and allophones in a second language: Evidence for distribution-based learning. *Studies in Second Language Acquisition*, 32(4): 581–606. https://doi.org/10.1017/S0272263110000276

Siew, C. S. Q. & Vitevitch, M. S. (2020) An investigation of network growth principles in the phonological language network. *Journal of Experimental Psychology: General*, 149(12): 2376–2394. https://doi.org/10.1037/xge0000876

Tabachnick, B. G. & Fidell, L. S. (2013) *Using multivariate statistics* (6th ed.). Upper Saddle River: Pearson Education.

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

Trofimovich, P., Kennedy, S. & Blanchet, J. (2017) Development of second language French oral skills in an instructed setting: A focus on speech ratings. *Canadian Journal of Applied Linguistics*, 20(2): 32–50. https://doi.org/10.7202/1042675ar

Wagner, R. K. & Torgesen, J. K. (1987) The nature of phonological processing and its causal role in the acquisition of reading skills. *Psychological Bulletin*, 101(2): 192–212. https://doi.org/10.1037/0033-2909.101.2.192

Zeigler, B. P., Muzy, A. & Kofman, E. (2018) *Theory of modeling and simulation: Discrete event and iterative system computational foundations* (3rd ed.). London: Academic Press.

Zhang, R. & Yuan, Z. (2020) Examining the effects of explicit pronunciation instruction on the development of L2 pronunciation. *Studies in Second Language Acquisition*, 42(4): 905–918. https://doi.org/10.1017/S0272263120000121

About the authors

Dennis Foung is a writing teacher at the School of Journalism, Writing and Media at the University of British Columbia. He has a keen interest in computer-assisted language learning and learning analytics.

Lucas Kohnke is a senior lecturer at the Department of English Language Education at the Education University of Hong Kong. His research interests include technology-supported teaching and learning, professional development using information communication technologies, and second language learning/acquisition.

Author ORCiD. Dennis Foung, https://orcid.org/0000-0002-6769-0582
Author ORCiD. Lucas Kohnke, https://orcid.org/0000-0001-6717-5719

https://doi.org/10.1017/S0958344023000071 Published online by Cambridge University Press