Language Teaching (2024), 1–9
doi:10.1017/S0261444824000120

CAMBRIDGE UNIVERSITY PRESS

REPLICATION RESEARCH

Replication research in the domain of perceived L2 fluency: Approximate and close replications of Kormos and Dénes (2004) and Rossiter (2009)

Viktoria Magne

University of West London, London, UK
Email: viktoria.magne@uwl.ac.uk

(Received 7 August 2023; revised 12 March 2024; accepted 29 March 2024)

Abstract

The primary objective of this paper is to contribute to the advancement of second language (L2) fluency research by outlining a specific proposal for future replication studies. The overarching goal is to assess the generalisability of the original findings of the two influential studies in the area of perceived fluency: Kormos and Dénes (2004) and Rossiter (2009). This objective will be achieved by first introducing the concept of L2 fluency that often conflates two categories: (1) overall language proficiency; (2) temporal features of speech production. The paper then highlights limitations in the current fluency research paradigm emphasising the variability in the methods employed for speech analysis and rating data collection. This diversity makes it somewhat challenging to compare results across various studies. In response to these challenges, the second part of the paper proposes several close and approximate replications of the two studies.

1. Introduction

Replication studies play a crucial role in scientific research by validating, refining, reinforcing and sometimes limiting the findings of initial studies (McManus, 2023). They contribute to the reliability, transparency and robustness of research, ensuring that results are reliable and generalisable (Marsden et al., 2018). Moreover, replication studies help identify the boundaries and limitations of existing research, contributing to the cumulative knowledge in the field (Porte & McManus, 2019). At the time of writing, in the context of second language (L2) fluency research, there is only one paper – by Gam and Ma (2023), who carried out a conceptual replication of Tavakoli and Hunter's (2018) study – which investigated L2 teachers' understanding of speech fluency. The scarcity of replication studies in the field of L2 fluency is alarming, given the considerable variability in the conceptualisation and operationalisation of L2 fluency in second language acquisition (SLA) research. Therefore, it is essential to undertake replications to gain a more profound understanding of the role of methodological variables and their effect on the robustness of the existing findings. This paper suggests replicating two influential correlational studies in the domain of perceived fluency by Kormos and Dénes (2004) and Rossiter (2009).

Fluency is a multifaceted phenomenon that has been variously defined in the literature. It is hard to find an agreed-upon definition of fluency as it spans from encompassing overall language proficiency to focusing on temporal phenomena, measured through the analysis of speech speed, pauses and repair (Tavakoli & Hunter, 2018). Lennon (1990) distinguished between narrow and broad definitions of fluency. In its broad sense, fluency is conceptualised as the 'spoken command of a foreign language' (Lennon, 1990, p. 389). In its narrow sense, fluency is seen as a separate, distinct and measurable component of oral proficiency. Further developing our understanding of fluency, Segalowitz (2016)

proposed a tripartite system that divides fluency into three distinct categories: L2 utterance, cognitive and perceived fluency. L2 utterance fluency is defined as a set of measurable temporal features that consist of speed, breakdown and repair measures (Tavakoli & Skehan, 2005). L2 cognitive fluency concerns the cognitive processes that underlie L2 utterances and is defined as the rapid mobilisation and retrieval of those processes (Segalowitz, 2016). L2 perceived fluency is the subjective evaluations of L2 oral performance by listeners expressed in numerical rating scales (e.g., Saito et al., 2018). Previous studies (e.g., Prefontaine et al., 2016) suggest there is a correlation between subjective perceived fluency scores and a set of objective utterance fluency measures. However, the precise elements of utterance fluency that influence perceived fluency scores still require additional investigation and validation.

1.1 Measuring perceived and utterance fluency

To measure utterance fluency, collected speech samples (normally gathered through a range of oral tasks) are transcribed for syntactic and acoustic analyses. For the syntactic analysis, one commonly used approach is to employ the Analysis of Speech Unit (AS-unit) (Foster et al., 2000). In the examples provided taken from Foster et al. (2000), an AS-unit boundary is denoted by an upright slash, and a clause boundary is indicated by a double colon. The Analysis of Speech Unit divides a single speaker's utterance into syntactic units that consist of an independent clause such as | Turn left |, or a subclausal unit, for instance, | how long you stay here |, accompanied by any subordinate clause(s) related to it (pp. 365–366). For instance, | It is my hope :: to study crop protections | represents two clauses and one AS-unit (p. 366). This type of analysis allows for the inclusion of units of speech segmentation which are common in naturally occurring speech.

The next step is the acoustic analysis, which includes different acoustic dimensions designed to capture parameters of spoken utterances. The acoustic dimensions utilised extensively and deemed robust in prior studies (Tavakoli & Skehan, 2005), namely breakdown, speed and repair fluency, constitute a framework for measuring oral fluency, referred to by Segalowitz (2016) as 'utterance fluency'. Speed fluency measures encompass aspects of speech indicating the speed of delivery, while breakdown measures signify disruptions or pauses, and repair measures are illustrated by instances of repetitions and reformulations during the speech production process (Tavakoli & Hunter, 2018). This tripartite framework allows researchers to operationalise the multifaceted construct of oral or 'utterance fluency' by dividing it into measurable sub-constructs. Quantifying the acoustic dimensions of fluency not only provides tangible metrics, reducing subjectivity in assessments, but also allows for the pinpointing of specific aspects of L2 speech that may contribute to disfluency. Additionally, it offers a better understanding of the cognitive aspects involved in L2 production.

Nevertheless, there remains CONSIDERABLE variability in the ways these three dimensions are calculated across studies (Bosker et al., 2013; De Jong, 2018; Suzuki & Kormos, 2020), meaning future researchers may have difficulty generalising from their findings. Typically, to calculate breakdown fluency, the number of filled and unfilled pauses within and between clauses are counted and divided by the length of speech. However, there is emerging evidence about the potential multidimensionality of breakdown fluency (Suzuki & Kormos, 2020). Several studies have reported that pause frequency, duration and location may affect L2 fluency judgements (Kormos, 2006; Saito et al., 2018), but the relative contribution of each of these dimensions to the perceived fluency scores requires further investigation. Speed fluency is estimated by dividing the total phonation time by the articulation rate and is expressed in the mean number of syllables per second. Repair fluency is usually computed by dividing the number of repetitions and self-corrections by the total number of words. The coding of utterance fluency across different dimensions should be performed by several trained researchers to ensure comparable levels of inter-rater agreement.

Listeners or raters also vary across studies on several dimensions, which, in turn, could affect the results. First, earlier studies in L2 fluency research employed listeners who are native speakers of English (Derwing et al., 2004; Riggenbach, 1991). However, given the position of English as the world's international language, the use of English as a common communication tool between L2 speakers is

https://doi.org/10.1017/S0261444824000120 Published online by Cambridge University Press

more likely (Pennycook, 2017). Therefore, to increase the applicability of the findings of fluency research to wider contexts, it is essential to explore L2 users' perception of L2 fluency (Magne et al., 2019; Rossiter, 2009). Second, most previous studies have included trained native speakers as raters (Cucchiarini et al., 2002; Lennon, 1990; Riggenbach, 1991); however, Prefontaine et al. (2016) found that there is no difference between the scores assigned by trained teachers and naïve listeners. Hence, further exploring the views of non-trained or naïve listeners will result in increased ecological validity of fluency studies. Another rater factor difference is whether to provide a pre-rating training session. Some researchers include listener training (Saito et al., 2018; Tavakoli et al., 2016) while some do not (Suzuki & Kormos, 2020) – which, in turn, may affect the features of spoken language raters focus on. Therefore, there has recently been a call to include a post-rating debrief session to gain insights into raters' understating of fluency and to collect supplementary qualitative data (Suzuki et al., 2021).

As seen from the brief overview of the literature, there is considerable methodological variability in perceived fluency research, which, in turn, colours researchers' interpretations of the outcomes. Replication studies will therefore enable researchers to check the validity of the existing constructs, to test out the existing methodological choices and to verify conclusions of the existing studies. Replication involves the repetition of a study's methodology and design, with or without alterations, and subsequent systematic comparison to enhance comprehension of the nature, repeatability and generalisability of its findings (McManus, 2023). The primary objective of a replication study is to test the validity of constructs and instruments, as well as to identify possible sources of variation in the results if consistent findings are not obtained. In essence, checking the validity of existing constructs through replication involves systematically reproducing the original study to see if the same theoretical constructs are consistently supported. An important distinction to note between an extension study and a replication study is that the primary aim of the extension study is not to critically question, revisit or reconsider the methodology of the previous study whereas, for the replication study, a comparison with the original study is essential, making it an integral part of the replication process (Porte & McManus, 2019). Replication is thus defined as a methodological tool that entails repeating a study systematically with the explicit aim of enhancing understanding and/or confirming the results of a previous study (McManus, 2023). The following section presents two studies that looked at perceived and utterance fluency in L2 speech using a mixed methods approach. These studies were selected because they used similar methods of data collection yet arrived at slightly different results.

2. Replication study 1: Kormos and Dénes (2004)

2.1 Background to the study

The first study suggested for replication investigated variables that predicted teachers' perceptions of oral fluency. The study also attempted to identify what features distinguished between fluent and non-fluent L2 speakers of English in the context of a Hungarian university. Kormos and Dénes' paper attempted to isolate the exact measures that can be utilised to help separate fluency from overall language proficiency. Kormos and Dénes (2004) is worthy of replication because of its considerable influence in the field (978 citations on Google Scholar on 6 January 2024), particularly in the area of listeners' perceptions of fluency and their associations with temporal features of L2 speech. In addition to the number of citations, Kormos and Dénes' study is a small-scale correlation study that could benefit from untangling certain multicollinearity issues that might have created uncertainty in the results.

To establish variables that differentiate fluent from non-fluent language learners and to determine the exact features that underpin raters' fluency judgements, Kormos and Dénes recruited 16 participants from a university in Hungary. The participants were equally divided into two groups based on their proficiency scores in the school English as a foreign language exam. The groups were labelled

https://doi.org/10.1017/S0261444824000120 Published online by Cambridge University Press

'advanced' and 'low-intermediate'. The advanced group was all female and the low-intermediate group was mixed (male and female) with ages ranging between 19 and 30 in both groups. Following two minutes of preparation time, all participants were asked to produce spontaneous speech describing a cartoon strip of their choosing. Performance was audio-recorded and later transcribed using a programme called *Transcriber* for the following ten temporal variables: speech rate, articulation rate, phonation-time ratio, mean length of runs, the number of silent pauses per minute, the mean length of pauses, the number of filled pauses, the number of disfluencies per minute, pace and space (i.e., the number of stressed words divided by the total number of words). In addition, the lexical diversity and the quantity of talk were also assessed. Three first language (L1)- and L2-speaking teachers were tasked with listening to the recordings and assessing the participants' performance using a five-point semantic differential scale (1 = least fluent, 5 = most fluent). They were also asked to provide qualitative comments on the assigned scores. Correlational analyses were conducted to determine variables that had affected the ratings. The Mann-Whitney U-test was used to establish if there was a difference between the fluent and non-fluent groups. The teachers' comments were summarised to determine the aetiology of the scores.

The results revealed that a set of variables appears to predict fluency scores assigned by L1- and L2-speaking teachers. These variables include speech rate, the mean length of run, phonation-time ratio and pace – with pace being a novel finding at the time. Based on the results, the authors argued that the number of stressed words per minute (i.e., pace) may be a better predictor of fluency than the number of syllables per minute. Importantly, Kormos and Dénes's study showed that the conceptualisation of perceived fluency by the listeners went beyond a merely temporal phenomenon, but included accuracy, lexical diversity, grammatical complexity and intonation, with accuracy particularly emphasised in the raters' qualitative comments. Those findings led the authors to suggest that accuracy should be considered for inclusion as one of the variables when investigating L2 fluency. Interestingly, a significant difference was not observed between the perceived fluency scores assigned by L1 and L2 teachers.

2.2 Approaches to replication

A close replication of Kormos and Dénes (2004) is suggested, which is an approach to replication wherein only one major variable is modified (Porte & McManus, 2019). The purpose of the close replication of this study is to advance our understanding of the robustness of the utterance fluency constructs (speed, breakdown and repair fluency) discussed earlier, and to verify whether possible issues of multicollinearity of fluency measures could account for the lack of significant findings for repair and breakdown fluency in the original study. This approach aims to investigate how intentionally altering a single variable enhances the findings of a previous study, thereby strengthening disciplinary knowledge (McManus, 2023; Porte & McManus, 2019).

Kormos and Dénes used 16 spontaneous speech samples produced by eight advanced and eight low-intermediate proficiency Hungarian speakers of English. First of all, the sample size of 16 is not sufficient to draw definitive conclusions. For example, given that the correlation between utterance and perceived fluency is large (Suzuki et al., 2021), and assuming the independently constructed predictors for speed, breakdown and repair, the G*Power analysis that helps estimate an approximate number of participants needed to obtain reliable results in terms of statistical power (Faul et al., 2007) puts the required sample size at 54. By increasing the number of speech samples, a replication study could verify the results of the original study that may have been affected by a Type II error, which is more likely to occur when a sample size is too small. Second, supplementing a correlation analysis with multiple regression analyses to extricate the relative weights of utterance fluency measures in assigned perceived fluency scores would help to avoid the issues of multicollinearity that may have affected the results of the original study. The original study found higher correlation coefficients for speed fluency but not breakdown fluency or repair fluency, which could be attributed to the large number of acoustic measures that were potentially interrelated (e.g., speech rate and mean length

https://doi.org/10.1017/S0261444824000120 Published online by Cambridge University Press

of run). For example, on closer examination of Table 3 in the original paper, it is evident that speech rate and phonation time are highly correlated. This is problematic as composite measures, such as speech rate, can incorporate various aspects of utterance fluency. This results in strong correlations with perceived fluency scores, but it lacks clarity regarding which constituent element of the composite measures has contributed to these correlations (Suzuki et al., 2021). As De Jong (2018) and Bosker et al. (2013) pointed out, studies that attempt to identify the relative contribution of the objective measures of fluency without considering the problem of multicollinearity should be interpreted with caution as they can lead to confounding results. Consequently, a replication of this kind could bring us closer to understanding possible confounding factors in the measures of fluency employed in the original study.

Another close replication would be to modify the speaking task itself. In the limitations section of the paper, Kormos and Dénes suggest that their study should be repeated with other types of tasks. Kormos and Dénes used a narrative task comprising of a cartoon strip with 6–10 thematically linked images. The participants were given two minutes of planning time to make sense of the story and to come up with a narrative around it. Such an approach works well with more proficient L2 speakers (e.g., Derwing & Munro, 2009); however, for beginner-level learners, using separate pictures with keywords seems more appropriate (see Saito & Shintani, 2016; Saito et al., 2016). A close replication would keep the rating task and analysis constant; however, it would employ a different task of speech elicitation. Instead of using a cartoon strip, a number of separate pictures with keywords would be employed to allow for a wider range of L2 fluency to be included (e.g., Saito et al., 2018). The potential contribution of such modification would clarify the role of the task in perceived fluency research. For example, Suzuki and Kormos (2023) have found that the constructs of speed and repair fluency vary across task types whereas breakdown fluency remains constant. In other words, running a replication study with different speaking tasks would further test the validity of the three constructs of fluency (speed, breakdown and repair fluency) as predictors of perceived fluency and their stability across task types.

An additional important point to consider in future re-visits to this and other similar papers is an improved clarity of qualitative data (Porte & Richards, 2012; Sato, 2020). In their original paper, Kormos and Dénes provide very little information about the way qualitative comments were collected and analysed. In the procedure description, the authors briefly mention that the raters were asked to comment on the scores for each participant with no further details regarding, for instance, the word limit or writing prompts for the task. Furthermore, Kormos and Dénes did not provide any explanations of how the data was coded or analysed. Echoing Sato's (2020) observations, I would argue that coding and analysis procedures should be an integral part of high-quality qualitative data analysis. Such findings could provide a valuable insight into participants' understanding of the task and the features they focused on when making fluency judgements. Building on Isaacs and Trofimovich's (2012) work, the themes could be then transformed into quantitative data by calculating word frequency for each theme (see Magne et al., 2019), thus further ensuring the reliability of the findings.

3. Replication study 2: Rossiter (2009)

3.1 Background to the study

The second study suggested for replication aimed to explore perceptions of L2 fluency by native and non-native speakers of English and was published in *The Canadian Modern Language Review* in 2009. This highly cited study (257 citations on Google Scholar on 6 January 2024) deserves replication because it was the first of its kind to include a group exclusively composed of L2 speakers. Prior to Rossiter's study, previous research in perceived fluency had focused almost exclusively on employing L1 speakers of English, more specifically teachers, as raters (Skehan, 2003). Despite its novelty at the time, the study remains small-scale, employing composite measures of fluency that pose challenges in

https://doi.org/10.1017/S0261444824000120 Published online by Cambridge University Press

interpreting the results as it remains unclear which temporal features each composite measure represents (Suzuki et al., 2021). Given this, a number of replications would be useful to test the robustness of the original findings and to advance knowledge and understanding of the ways native and non-native speakers perceive L2 fluency.

To explore fluency development over time (Time 1 and Time 2), Rossiter recruited 24 adult English Second Language (ESL) learners (11 men, 13 women) as speakers and three groups of listeners. The first group of listeners was labelled the 'expert' group, which consisted of six experienced native speaker (NS) ESL teachers (three women, three men) with ten years of experience of teaching ESL and phonology specifically. The 'novice' group included 15 inexperienced NS listeners studying for an education degree. The third group of listeners consisted of 15 advanced non-native speakers (NNS) enrolled in a translation degree. The talkers had intermediate proficiency in English with varied educational backgrounds. The age of the speakers ranged from 21 to 59 (M = 35 years) with an average of three years and seven months of experience in English-speaking Canada. Each participant was given the same eight-frame picture description task with one minute of planning. The picture descriptions were audiotaped at Time 1 (T1) and ten weeks later at Time 2 (T2). The voice samples ranged from 1.4 to 9.1 min ($M = 3.7$ min). The collected samples (at T1 and T2) were coded for self-repetition, self-correction, false starts, reformulation, asides and unfilled pauses of 400 msec or longer to be subtracted from the total number of syllables and divided by the total number of seconds to produce a measure of pruned syllables per second. Speech rate and mean length of run were also calculated. The voice clips were then randomly paired across time and then randomised across speakers and presented to the three groups of listeners.

The raters were instructed to listen to each sample, write down their first impressions and provide a rating for each using a nine-point Likert scale (1 = extremely dysfluent, 9 = very fluent). They were also instructed to assign a different rating to each member of a paired speech sample. Quantitative and qualitative analyses were carried out. The quantitative analyses included a repeated measures one-way ANOVA and Pearson's correlations. The qualitative results were separated into positive or negative at T1 and T2. Only the negative comments were analysed qualitatively as they were the majority.

The results revealed that the NNS group of listeners gave the lowest ratings at T1 and T2 than either of the NS groups. However, statistically significant results were only observed between Novice NS and NNS groups with the novice group giving the highest ratings. There was no statistically significant difference between T1 and T2 scores, which means that learners were not perceived to have improved over ten weeks of study. In terms of correlates of fluency, the following were found statistically significant: pruned syllables per second correlated with higher ratings by listeners; increased pausing, on the other hand, correlated with lower ratings of speaking fluency, which is a consistent finding in the literature. The quantity of the negative comments varied across groups with the smallest number recorded by the NNS group. The qualitative comments pointed to pausing, self-repetition, speech rate and the use of non-lexical fillers as major contributors to the ratings. Despite the fact that the listeners were asked to focus on temporal aspects of the oral productions, approximately one-quarter of the negative impressions they recorded were classified as non-temporal, relating to pronunciation, grammar and vocabulary.

3.2 Approaches to replication

Similarly to Kormos and Dénes's (2004) study, the sample size is an issue that may have affected the results of the original study. G*Power analysis (Faul et al., 2007) indicates that in order to at least maintain medium statistical power with Rossiter's (2009) design, 150 participants are required (50 people in each group). With an increased sample size, a possible close replication would focus on the utterance fluency measures employed in the study. A replication study would employ the same task type, the same data collection procedure and three groups of listeners (expert native, novice native and non-native); however, the approach to coding fluency would change. The original study employed the composite measures of fluency, such as speech rate and mean of length of run; however, they may not have

https://doi.org/10.1017/S0261444824000120 Published online by Cambridge University Press

been appropriate as they are difficult to interpret when it comes to disentangling the utterance from the perceived fluency. As Suzuki et al. (2021) pointed out, it would be particularly difficult to pinpoint which exact features would in fact predict the obtained scores and could lead to findings that are hard to interpret as the weighting of each temporal feature is unclear, which is what happened in the original study. Moreover, there is mounting evidence to suggest that not only pause duration but also pause location is important in predicting perceived fluency results, therefore it becomes difficult to generalise the original study's findings given the nature of the fluency measures employed. A possible way forward in terms of replication would be to code the speech samples for the well-established measures of breakdown, speed and pause fluency and use those measures to determine their individual predictive power. The potential outcome of the coding modification would be a better understanding of the original study's findings, especially in terms of individual contributing factors of each aspect of utterance fluency to the raters' scores vis-à-vis their assigned group (expert native vs non-native vs novice native). For example, Saito et al. (2018) found that one of the predictors of native speakers' judgements of L2 fluency was final-clause pause ratio, whereas L2 speakers' perceived fluency was predicted by mid-clause pause ratio and excluded final-clause pauses (Magne et al., 2019). Replication with the proposed modifications will shed light on whether the differences with more recent studies are due to the conflating of the dimensions of utterance fluency measures in the original study.

Running a stepwise multiple regression analysis instead of the set of correlation analyses employed in the original study would allow to control for the inter-collinearity of utterance fluency measures (Magne et al., 2019; Suzuki et al., 2021) to find the best regression model that accounts for the highest percentage of the total variance. However, there are limitations associated with a stepwise multiple regression analysis, such as the danger of fitting the data into a model that best suits the desired result. A better approach would be to include the variables that are guided by theory rather than trying to achieve the 'best' result. The use of mixed effects modelling has been suggested as a way to obtain more accurate results and to deal with more complex data as it permits researchers to explore not only between-subject variables but also within-subject variables (Suzuki et al., 2021).

The next step would entail an approximate replication that would focus on the non-native listeners and involve changes to the context of the study, as approximate replication allows for two variables to be modified (Porte & McManus, 2019). The original study was conducted in western Canada where English is the main language of communication. As a result, the participants – who came from a variety of linguistic backgrounds – were already more familiar with several different varieties of L2 English, which, in turn, may have affected what specific features they paid attention to when making the judgments. It would therefore be interesting to see whether findings would hold in an English as a foreign language (EFL) context, where English is not used as the main language of communication. Consequently, employing a homogenous group of listeners with no immersion experience may affect the results. Homogenous listeners (raters) may be particularly sensitive to articulation rate when they make fluency judgements due to their exposure to speech samples produced by native speakers whose speech is controlled for speed in the teaching materials they receive. Some listening comprehension studies have also pointed out that L2 listeners' experiences with specific linguistic varieties have a significant effect on perceptual representation of L2 speech (Atagi & Bent, 2016).

A note on the qualitative element of the original study: While the possibility of replicating a qualitative study remains a contested issue (Porte & McManus, 2019), it is important to point out that in mixed-method studies with a strong quantitative focus, reporting qualitative data becomes almost an afterthought, given their perceived lack of rigour and systematicity. A way forward would be to employ a more structured analysis, such as interpretive phenomenological analysis (Smith et al., 2009) or thematic analysis (Braun & Clarke, 2006), that follows a set protocol for coding, analysis and reporting.

4. Conclusion

Oral fluency is extremely important in language teaching and is often a marker of language proficiency. Previous research studies in the area of fluency are in agreement about the complex nature

https://doi.org/10.1017/S0261444824000120 Published online by Cambridge University Press

of fluency and the ways of measuring it. While fluency comprises multiple dimensions, the focus of this paper has been on the relationship between perceived and utterance fluency. More specifically, the exact link between subjective perceived fluency judgements provided by raters and objective measures of fluency are still not diffidently explored. Therefore, replication studies can assist in testing the identified measures of fluency in a more systematic way. The present paper has made a number of specific methodological suggestions for replicating the studies by Kormos and Dénes (2004) and Rossiter (2009). Both recommended studies attempted to pinpoint the exact measures of fluency that have affected the ratings assigned by listeners. Kormos and Dénes (2004) moved the field forward by looking at different composite measures of fluency, while Rossiter (2009) made a significant contribution to fluency research by extending it to second language speakers. These studies are good candidates for replication, as the original data is available from the researchers and the research instruments are well-known and widely used. This paper has suggested a number of close and approximate replications that would address the shortcomings of the original studies and enhance our current understanding of perceived and utterance fluency.

Acknowledgements. The author wishes to thank Shungo Suzuki (Waseda University) and Yui Suzukida (Juntendo University) for their helpful feedback on the earlier versions of the manuscript.

References

Atagi, E., & Bent, T. (2016). Auditory free classification of native and nonnative speech by nonnative listeners. *Applied Psycholinguistics*, *37*(2), 241–263. doi:10.1017/S014271641400054X

Bosker, H. R., Pinget, A.-F., Quené, H., Sanders, T., & de Jong, N. H. (2013). What makes speech sound fluent? The contributions of pauses, speed and repairs. *Language Testing*, *30*(2), 159–175. doi:10.1177/0265532212455394

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. doi:10.1191/1478088706qp063oa

Cucchiarini, C., Strik, H., & Boves, L. (2002). Quantitative assessment of second language learners' fluency: Comparisons between read and spontaneous speech. *Journal of the Acoustical Society of America*, *111*(6), 2862–2873. doi:10.1121/1.1471894

De Jong, N. (2018). Fluency in second language testing: Insights from different disciplines. *Language Assessment Quarterly*, *15*(3), 237–254. doi:10.1080/15434303.2018.1477780

Derwing, T. M., & Munro, M. J. (2009). Putting accent in its place: Rethinking obstacles to communication. *Language Teaching*, *42*(4), 476–490. doi:10.1017/S026144480800551X

Derwing, T. M., Rossiter, M. J., Munro, M. J., & Thomson, R. I. (2004). Second language fluency: Judgments on different tasks. *Language Learning*, *54*(4), 655–679. doi:10.1111/j.1467-9922.2004.00282.x

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175–191. doi:10.3758/BF03193146

Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. *Applied Linguistics*, *21*(3), 354–375. doi:10.1093/applin/21.3.354

Gan, Q., & Ma, L. (2023). Examining the perceptions and self-reported practices of L2 teachers in China regarding oral fluency: A conceptual replication and extension. *Language Teaching Research*. doi:10.1177/13621688231186857

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. doi:10.1017/S0272263112000150

Kormos, J. (2006). *Speech production and second language acquisition*. Lawrence Erlbaum Associates. doi:10.4324/9780203763964

Kormos, J., & Dénes, M. (2004). Exploring measures and perceptions of fluency in the speech of second language learners. *System*, *32*(2), 145–164. doi:10.1016/j.system.2004.01.001

Lennon, P. (1990). Investigating fluency in EFL: A quantitative approach. *Language Learning*, *40*(3), 387–417. doi:10.1111/j.1467-1770.1990.tb00669.x

Magne, V., Suzukida, Y., Ilkan, M., Tran, M., Suzuki, S., & Saito, K. (2019). Exploring the dynamic nature of second language listeners' perceived fluency: A mixed-methods approach. *TESOL Quarterly*, *53*(4), 1139–1150. doi:10.1002/tesq.528

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

McManus, K. (2023). How and why to conduct a replication study. In A. Mackey, & S. M. Gass (Eds.), *Current approaches in second language acquisition research: A practical guide* (pp. 334–351). Wiley. doi:10.31219/osf.io/x3e4u

Pennycook, A. (2017). *The cultural politics of English as an international language*. Routledge.

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

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

Prefontaine, Y., Kormos, J., & Johnson, D. E. (2016). How do utterance measures predict raters' perceptions of fluency in French as a second language? *Language Testing*, *33*(1), 53–73. doi:10.1177/0265532215579530

Riggenbach, H. (1991). Toward an understanding of fluency: A microanalysis of nonnative speaker conversations. *Discourse Processes*, *14*(4), 423–441. doi:10.1080/01638539109544795

Rossiter, M. J. (2009). Perceptions of L2 fluency by native and non-native speakers of English. *Canadian Modern Language Review*, *65*(3), 395–412. doi:10.3138/cmlr.65.3.395

Saito, K., Ilkan, M., Magne, V., Tran, M., & Suzuki, S. (2018). Acoustic characteristics and learner profiles of low, mid and high-level second language fluency. *Applied Psycholinguistics*, *39*(3), 593–617. doi:10.1017/S0142716417000571

Saito, K., & Shintani, N. (2016). Do native speakers of North American and Singapore English differentially perceive second language comprehensibility? *TESOL Quarterly*, *50*(2), 421–446. doi:10.1002/tesq.234

Saito, K., Trofimovich, P., & Isaacs, T. (2016). Second language speech production: Investigating linguistic correlates of comprehensibility and accentedness for learners at different ability levels. *Applied Psycholinguistics*, *37*(2), 217–240. doi:10.1017/s0142716414000502

Sato, M. (2020). Learner attitudes and attention to form in peer interaction: A proposal to replicate Adams et al. (2011) and Philp et al. (2010). *Language Teaching*, *55*(3), 407–416. doi:10.1017/S0261444820000610

Segalowitz, N. (2016). Second language fluency and its underlying cognitive and social determinants. *International Review of Applied Linguistics in Language Teaching*, *54*(2), 79–95. doi:10.1515/iral-2016-9991

Skehan, P. (2003). Task-based instruction. *Language Teaching*, *36*(1), 1–14. doi:10.1017/S026144480200188X

Smith, J. A., Flower, P., & Larkin, M. (2009). Interpretative phenomenological analysis: Theory, method and research. *Qualitative Research in Psychology*, *6*(4), 346–347. doi:10.1080/14780880903340091

Suzuki, S., & Kormos, J. (2020). Linguistic dimensions of comprehensibility and perceived fluency: An investigation of complexity, accuracy, and fluency in second language argumentative speech. *Studies in Second Language Acquisition*, *42*(1), 143–167. doi:10.1017/S0272263119000421

Suzuki, S., & Kormos, J. (2023). The multidimensionality of second language oral fluency: Interfacing cognitive fluency and utterance fluency. *Studies in Second Language Acquisition*, *45*(1), 38–64. doi:10.1017/S0272263121000899

Suzuki, S., Kormos, J., & Uchihara, T. (2021). The relationship between utterance and perceived fluency: A meta-analysis of correlational studies. *Modern Language Journal*, *105*(2), 435–463. doi:10.1111/modl.12706

Tavakoli, P., Campbell, C., & McCormack, J. (2016). Development of speech fluency over a short period of time: Effects of pedagogic intervention. *TESOL Quarterly*, *50*(2), 447–471. doi:10.1002/tesq.244

Tavakoli, P., & Hunter, A.-M. (2018). Is fluency being 'neglected' in the classroom? Teacher understanding of fluency and related classroom practices. *Language Teaching Research*, *22*(3), 330–349. doi:10.1177/1362168817708462

Tavakoli, P., & Skehan, P.. (2005). Strategic planning, task structure, and performance testing. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 239–273). John Benjamins Publishing Company. doi:10.1075/lllt.11.15tav

Viktoria Magne is an Associate Professor in Education Studies and Early Childhood at the University of West London, UK. Her main research interests are in the area of applied linguistics with a particular focus on sociolinguistic aspects of second language speech. She has also collaborated on projects related to second language acquisition, psycholinguistics and student assessment. Viktoria has published articles in a number of journals including *TESOL Quarterly*, *Studies in Second Language Acquisition* and *Applied Psycholinguistics*.

Cite this article: Magne, V. (2024). Replication research in the domain of perceived L2 fluency: Approximate and close replications of Kormos and Dénes (2004) and Rossiter (2009). *Language Teaching* 1–9. https://doi.org/10.1017/S0261444824000120