Studies in Second Language Acquisition (2025), 1–21
doi:10.1017/S0272263125101095

METHODS FORUM

Saving the reliability of inhibitory control measures? An extension of Huensch (2024) and Hui and Wu (2024)

Zhiyi Wu, Ruirui Jia and Bronson Hui

The Graduate Program of Second Language Acquisition, School of Languages, Literatures, and Cultures, University of Maryland, College Park, MD, USA
Corresponding author: Zhiyi Wu; Email: zhiyiw1@umd.edu

(Received 20 December 2024; Revised 28 May 2025; Accepted 17 June 2025)

Abstract

In a close replication study of Darcy et al., (2016), Huensch (2024) reported a lack of clear relationships between inhibitory control (IC) and phonological processing, contrary to the initial findings. Given the general unreliability of response-time differences, which are often the basis of IC measures and could potentially mask small effects, we performed secondary analyses on Huensch's (2024) open data set to investigate (a) the extent to which the reliability of IC measures could be improved using model-based approaches (Hui & Wu, 2024), (b) the correlations between the different IC tasks, and (c) their predictive power for phonological processing, based on the more reliable indices. Results showed that model-based approaches generally improved reliability, and particularly for the Stroop and Simon tasks to acceptable levels. Yet, correlations between IC tasks remained low, and partial correlation and hierarchical regression still failed to reveal significant relationships between IC and phonological processing, further confirming Huensch's (2024) findings.

Keywords: reliability; inhibitory control; phonological processing; replication; response time differences; model-based approaches

Introduction

Instrument reliability is often overlooked as manifested in nonreports in research papers (McKay & Plonsky, 2021). Recently, issues surrounding reliability have received increased attention in second language acquisition (SLA), especially in studies employing tasks that elicit response-time (RT) differences (e.g., Buffington et al., 2021). For some research involving cognitive aptitude, a great deal is at stake because of the heavy reliance on RT differences as an individual difference measure (e.g., larger or smaller RT differences indicating stronger ability for an individual). Indeed, investigations into inhibitory control (IC) and its relationships with external variables (e.g., proficiency or

phonological processing) would fall into this category due to the type of measurement of IC (e.g., Darcy et al., 2016; Huensch, 2024); therefore, it is important for researchers to understand the issues around instrument reliability more thoroughly and identify solutions to mitigate any limitations inherent to RT-difference measures. While efforts have been made to improve the reliability of RT differences (Hui & Wu, 2024), the extent of the problem and the potential consequences of relying on such measures in research specifically involving IC in SLA have not yet been fully documented and illustrated. In addition, the degree to which model-based approaches could improve the reliability of IC measures for second language (L2) research is not entirely clear. The implications of the potentially improved reliability on the predictive validity of the measure also remain conceptual. Thus, in this paper, we perform a secondary analysis on an open dataset, initially collected for the examination of the relationship between IC and phonological processing for L2 learners (Huensch, 2024), to fill these gaps. Our additional, broader aim is also to raise L2 researchers' awareness of the reliability issues surrounding measures of cognitive individual differences.

Reliability challenges in RT-based individual differences measures: The case of inhibitory control measures

Generally, the reliability of an instrument refers to the extent to which it consistently yields the same score when used under the same measuring conditions with the participants (Cohen et al., 2017). This fundamental measurement principle applies across all individual differences research, with particular implications for RT measures widely used in SLA and cognitive psychology (e.g., Buffington et al., 2021; Maie, 2022). Reliability in correlational/individual differences studies is typically based on classical test theory to indicate the extent to which an instrument can rank individuals consistently (Hedge et al., 2018). In general, a number of factors can undermine reliability: A reduction in variance between individuals while error variance remains constant, or an increase in error variance while the variance between participants stays the same (Hedge et al., 2018). In other words, two potential sources of low reliability are (a) high measurement errors and (b) low between-participant variation. Both of these issues affect numerous individual difference measures in SLA, with RT-difference measures being particularly vulnerable (Hui & Wu, 2024; McKay & Plonsky, 2021).

A high level of measurement error can lead to "non-systematic change between individuals" (Hedge et al., 2018, p. 1167) across testing sessions, thereby undermining reliability. Whether it is due to participant bias or item design within the instruments themselves, measurement errors can introduce variability in the data that is not related to the construct being measured, subsequently mitigating the validity of the results we observe. Additionally, reliability may suffer when there is little variation between participants or when the sample is particularly homogeneous (Hedge et al., 2018). This limitation is especially critical for correlational or individual differences research that "examines factors that distinguish between individuals within a population (i.e., between-subject variance)" (Hedge et al., 2018, p. 1166). To clarify this point further, when participants exhibit comparable performance on cognitive tasks, their resulting scores cluster together with insufficient differentiation between individuals. This restricted range of scores undermines the instrument's capacity to produce consistent rankings of individuals across multiple measurements. Even small measurement errors can alter participants' relative standings when the true between-participant differences are minimal. Consequently, correlational analyses using such measures may fail to detect genuine relationships not due to the absence of such relationships, but rather because the

https://doi.org/10.1017/S0272263125101095 Published online by Cambridge University Press

measurement instrument cannot reliably distinguish between participants' abilities. This methodological limitation is particularly problematic in individual differences research, where the goal is to precisely identify how variations in, for example, cognitive ability relate to learning outcomes.

When it comes to RT differences measures (i.e., a type of measure that depends on subtracting the RTs in different conditions), reliability can be low due to what Hedge et al. (2018) termed the "reliability paradox" (e.g., Tan & Yap, 2016; Buffington et al., 2021). These measures typically show robust effects at the group level but poor reliability for individual differences. This occurs potentially because subtracting condition means (e.g., incongruent minus congruent RTs) inherently reduces between-participant variance, a crucial component for reliability in individual differences research. Moreover, with many trials per participant, the magnitude of an individual's RT difference tends to show substantial within-person variability. Consequently, these measures struggle to consistently rank individuals across different subsets of trials, undermining their reliability for capturing individual differences in linguistic knowledge (e.g., Hui & Jia, 2024), processing (e.g., Frinsel & Christiansen, 2024), as well as cognitive ability, such as procedural memory capacity (e.g., Buffington et al., 2021) and IC (e.g., Huensch, 2024).

For RT-difference tasks that tap into IC, Hedge and colleagues (2018, Studies 1 and 2) examined the test-retest reliability of four response inhibition tasks (i.e., the Eriksen flanker task, Stroop task, go/no-go task, and the stop-signal task) commonly used in cognitive psychology and neuroscience. Calculating the intraclass correlation coefficient (ICC), the authors found that none of the four tasks met the reliability of .80, which was considered excellent, and that only two measures marginally met the threshold of being substantial (.60). In particular, the Stroop task (i.e., the task that requires participants to name the color of a presented word while inhibiting naming the word itself which could be a color word different in its ink color) had ICC values of .60 (session 1) and .67 (session 2), while the go/no-go task had an ICC of .76 in both sessions. Similarly, in Hedge et al. (2022), where researchers examined the reliability of multiple executive function tasks across several datasets, it was also reported that the reliability of behavioral measures, namely RT costs and error costs in conflict tasks, such as Flanker and Stroop tasks, is generally low. With the ICC of the RT costs ranging from .38 to .65 in the Flanker task and from .38 to .66 in the Stroop task (see the supplementary material A in Hedge et al., 2019), the RT costs over a four-week period indicate only a low to moderate test-retest reliability. Moreover, for the Simon task/Spatial Stroop task (i.e., participants were asked to name the color/meaning of a stimulus and ignore its location), despite its relatively higher test-retest reliability reported by Hedge et al. (2022), most of its ICCs still do not achieve the threshold of .80 across different datasets (i.e., dataset 1: .74; dataset 3: .60; dataset 5: .67; dataset 6: .72), displaying only a moderate level of reliability. Given the findings, researchers have argued that the unsatisfactory reliability could be attributed to the lack of between-participant variability in the data (Hedge et al., 2018). This idea is supported by the observation that the ICC for RT differences is generally lower than that for RTs within each component (i.e., congruent and incongruent conditions) (Hedge et al., 2022). This is because the calculation of RT differences involves subtracting the RT in one condition from the RT in another, which inherently reduces between-participant variability and, as a result, contaminates reliability (Hedge et al., 2018, 2022).

Given the reduced between-participant variance and potential measurement errors associated with RT measures—such as participants feeling fatigued when pressing buttons or being inattentive to stimuli—these findings (i.e., low reliability of many RT-based measures for IC and other cognitive constructs) should be alarming for any

https://doi.org/10.1017/S0272263125101095 Published online by Cambridge University Press

serious researchers using RT-difference measures as indicators of individual differences in SLA and beyond. This does not imply, however, that RT-difference measures should be abandoned for indexing individual differences (Hui & Wu, 2024). Rather, it highlights the importance of carefully considering how IC and, more generally, cognitive differences between individuals can be measured with high reliability. With more reliable measures, researchers could reveal more precise and meaningful insights into individual differences in relation to L2 learning and processing. This is especially important for instructed SLA researchers who often seek to understand which learners benefit most from particular intervention by exploring aptitude × treatment interactions. Such exploration not only reveals how individual differences and external factors work together to influence language learning outcomes but also provides insight into the underlying processes at play (DeKeyser, 2021), which are core foci of SLA research.

In addition to the theoretical rationale of examining RT differences, it should be noted that, from a statistical point of view, RT-difference measures indexing individual differences are often placed on the predictor side of a regression equation, where perfect reliability is assumed. When this assumption is violated, findings regarding the predictive validity of IC to external variables such as phonological processing could be undermined. To further illustrate exactly how unreliability can mask important relationships, we present a hypothetical example here, where the outcome (y) represents phonological processing and the predictor (x) is IC. For the sake of clarity in this example, we only vary the level of reliability on x and assume what should never be assumed: that phonological processing (y in Figure 1 below) is measured reliably (e.g., at .90). Several simulated scenarios are visualized in Figure 1, where the reliability of x (IC) decreases from 1.00 to .20 across the panels. The first panel to the left shows the true, positive, and strong relationship between the two variables. As the reliability of x decreases, the data points spread more widely on the x-axis due to the increased unreliability, which subsequently causes the slope of the regression line to flatten, and eventually, the slope might not be significantly different from zero (no relationship). In other words, the unreliability of IC measures, depending on the extent, can mask any meaningful relationships with phonological processing. It is also worth noting that the baseline here is a strong relationship (left panel). If the true relationship is moderate or weak, there is then an even slimmer chance that it can be observed with an unreliable x variable.

While this is a serious issue, there is already some awareness in the field. For example, Huensch (2024) rightly pointed out that correlational studies can lose power as a result of the lack of between-participant variability in the inhibition score, an underlying cause for the unreliability. To further refine our understanding of the reliability issues, it would be useful to lay bare the extent of the problem, for example, by examining how different IC tasks might correlate with each other (or not). Also, the field would benefit from knowing if there are any effective solutions to the problem, for example, by employing more contemporary, model-based approaches to estimate a more reliable IC score for an individual learner.

Although we focus on IC measures in this paper, the reliability challenges we have discussed extend to many RT-based individual differences measures in SLA and cognitive science more broadly. The "reliability paradox" affecting IC tasks similarly impacts many other RT-differences measures, such as those measuring lexical access (Hui et al., 2025; Zhang et al., 2025), syntactic processing (Fang & Wu, 2022), and working memory capacity (Unsworth & Engle, 2005). Understanding and addressing these measurement challenges is thus important not only for improving IC research

https://doi.org/10.1017/S0272263125101095 Published online by Cambridge University Press

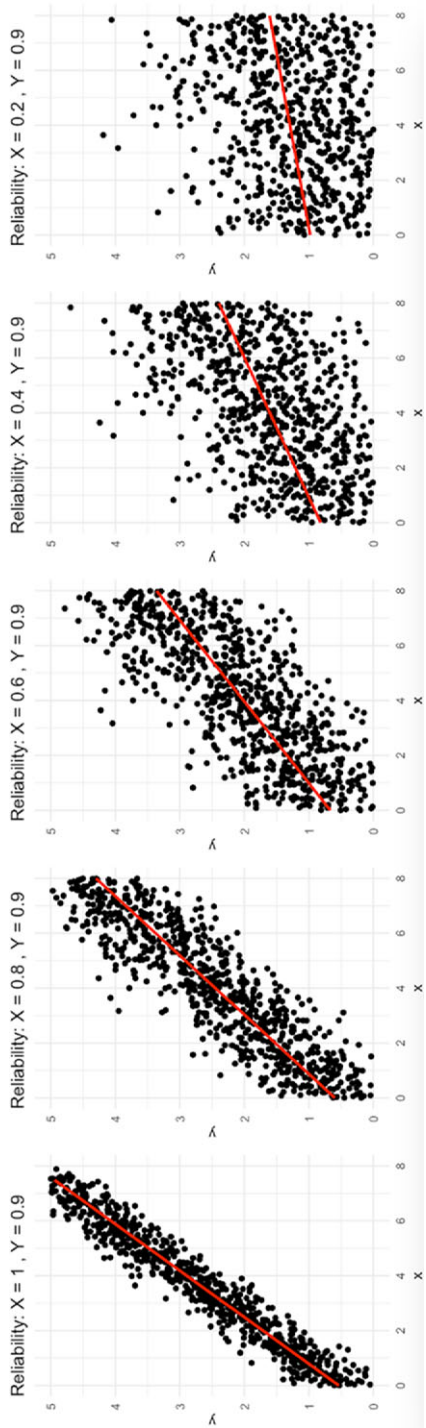

Figure 1. Visualization of the estimation change with various degrees of measurement errors involved.

https://doi.org/10.1017/S0272263125101095 Published online by Cambridge University Press

specifically but also for enhancing the methodological rigor of individual differences research across multiple domains in SLA.

Estimating the reliability of RT differences

For RT-difference measures (e.g., the Stroop task), researchers often follow classical test theory, aggregating trials into a single score by calculating the mean RT in the congruent and incongruent conditions and subtracting them for RT differences for further analyses. However, this approach often contaminates trial-by-trial variation, resulting in the reduction of not only reliability but also effect sizes and correlation (Rouder & Haaf, 2019). To address this issue, Rouder and Haaf (2019) advocated the application of mixed-effects models with trial-by-trial analysis. Taking into account both participant and item variability, the researchers reanalyzed the data shared by Hedge et al. (2018) and found a much-improved test-retest reliability for both the Stroop and the Flanker tasks. In SLA, Hui and Wu (2024) evaluated the effectiveness of model-based approaches by comparing three methods for estimating the reliability of RT differences: computations based on (a) raw RTs, (b) by-participant *z*-transformed RTs, and (c) the model-based estimation. The authors found that the model-based approach can outperform the other two methods in estimating reliability. However, this superiority is contingent upon certain conditions, namely moderate measurement error and a limited number of items. Despite the limitations, their findings underscore the potential of the model-based approach as a promising alternative to estimating the reliability of RT-difference tasks in SLA, potentially enhancing the validity of studies that rely on RT-difference measures.

Despite its advantages, not many studies to date have adopted the model-based approach for reliability estimation, and the degree to which the model-based approach could improve reliability in L2 research is still not clear. Moreover, whether the improved reliability derived from the model-based approach could actually influence the predictive power of a measure still lacks sufficient empirical evidence. Thus, the current study aims to not only further the investigation of how the model-based approach could improve reliability but also examine whether the improved reliability could lead to more accurate prediction of an external measure. To achieve this, we situated our study in the context of the relationship between inhibition control and L2 phonological processing and conducted a secondary analysis using an open-source database shared by Huensch (2024).

Huensch (2024) on inhibitory control and phonological processing

Before proceeding to describe the present study, we provide a brief review of Huensch (2024) to contextualize our secondary analysis. We chose this study because the author has commendably made the data publicly available (Huensch, 2022), for which we are truly thankful. After all, this sharing has enabled the current methodological investigation. The author carried out a preregistered, close replication study of Darcy et al. (2016), examining the relationships between IC and phonological processing. Huensch (2024) included two additional IC tasks (i.e., a Stroop task and a Simon task, both described above), aside from the retrieval-induced inhibition task originally included in Darcy et al. (2016). The inclusion of these two additional tasks is because these are "classic test[s] of prepotent response inhibition" (Huensch, 2024, p. 3), measuring intentional, behavioral resistance to an immediately distracting stimulus, in contrast to

https://doi.org/10.1017/S0272263125101095 Published online by Cambridge University Press

the retrieval-induced inhibition task which "represents unintentional, cognitive, resistance to proactive interference" (Huensch, 2024, p. 2). That is to say, these two tasks allow researchers to capture additional, and perhaps different, aspects of IC that are hypothesized to have a stronger relationship with production skills but that had not been examined in the initial study. The initial retrieval-induced inhibition task contained three phases: During the first phase, participants were instructed to memorize 18 words grouped into three different categories. In the second phase, they were prompted to recall three words from two of the three categories, and finally, in the third phase, participants were asked to identify whether the words presented had appeared in the first phase. Assessment of IC involved three trial conditions: (a) the practiced condition, (b) the inhibited condition (i.e., words not practiced but belonging to a practiced category), and (c) the controlled condition (i.e., words not practiced), with the latter two being the critical conditions. The retrieval-induced inhibitory score was calculated by dividing the median RT for the items in the inhibited condition by the median RT for the items in the control condition. A score greater than 1 indicates greater IC, with a higher value reflecting stronger inhibition (Huensch, 2024). As for the Simon task, each participant's Simon score was computed by subtracting the mean RT for the congruent items (when the location of the text stimulus on the screen matches that of the key on the keyboard to be pressed) from the mean RT for the incongruent items. Similarly, the Stroop score was determined for each participant by subtracting their mean RT for the neutral items (when the stimulus was a string of symbols) from their mean RT for the incongruent items (when the stimulus was text and did not match the color of it). For both the Simon and the Stroop tasks, lower scores indicate better IC as they reflect faster responses to incongruent items (Huensch, 2024). All three scores were calculated based on correct responses only, with incorrect responses being excluded during the data preprocessing stage. For phonological processing, Huensch (2024) followed the operationalization of Darcy et al. (2016). L2 learners' phonological perception was assessed by the speeded ABX categorization task with Spanish vowel and consonant contrasts /e-e̞j/ and /d-ɾ/ as critical experimental items, while their production was measured by the delayed sentence repetition task involving the same contrasts.

Data collected from 58 participants did not replicate the findings of the initial study. Specifically, the Spearman partial correlation analyses demonstrated no statistically significant relations between retrieval-induced inhibition and vowel/consonant perception and production, and the hierarchical regression revealed that inhibition was not a significant predictor of vowel perception accuracy. In terms of the additional IC tasks, neither the Stroop task nor the Simon task displayed a clear relationship with vowel perception and production, nor with consonant perception and production, similar to the retrieval-induced inhibition task. Overall, the findings from Huensch (2024) suggest that "no strong, clear, or consistent relationship emerges between inhibitory control and L2 perception/production skills" (Huensch, 2024, p. 17).

Huensch (2024) argued that the discrepancies in the results may be related to the possibilities that (1) the relationship between IC and L2 phonological processing might be weak, if not null; (2) the inhibition tasks used may not effectively capture individual differences in inhibition; and (3) variations in study features could potentially influence the results. Regardless of the reasons, methodologically, the author acknowledged the challenges for reliably measuring IC due to limited between-participant variability, and this lack of variability could, in turn, reduce statistical power, resulting in the null results that were observed.

https://doi.org/10.1017/S0272263125101095 Published online by Cambridge University Press

Huensch's insights, combined with the model-based approaches tested by Hui and Wu (2024) and Rouder and Haaf (2019), motivated applying a model-based approach to the data in Huensch (2024), which could potentially mitigate the limitations in the unreliability of IC tasks. This approach may offer the promise of more precise reliability estimates for these tasks and better accounts for the inherent variability in RT-based measures. If successful, it can strengthen the case for a weak, if not null, relationship between IC and phonological processing. Moreover, Huensch's (2022) dataset also offers a great opportunity to reexamine the correlation between different IC measures. This investigation can potentially tease out the confounding statistical consideration (i.e., lack of between-participant variability) regarding the low correlation observed in the literature (Hedge et al., 2018; Rey-Mermet et al., 2018) and unveil important questions regarding whether inhibition in different tasks is a unified concept or not (Rouder & Haaf, 2019). Lastly, as has been mentioned earlier, the potential for enhancing the predictive power of these inhibition tasks through improved reliability has not been well supported by empirical evidence. In other words, there is a lack of robust evidence demonstrating whether these tasks would more accurately predict L2 phonological processing if their reliability were enhanced. This gap leaves uncertainty regarding the extent to which boosting reliability could lead to better predictive outcomes for these tasks.

Thus, the current study aims to extend the work of both Huensch (2024) and Hui and Wu (2024) by applying the model-based approach to the RT data shared in Huensch (2022). By doing so, we hope to provide a more precise estimation of the reliability of these tasks, thereby contributing to a clearer understanding of the relationship between not only different IC measures but also the relationship between IC and L2 speech processing.

The present study

Building upon the work of Huensch (2024) and Hui and Wu (2024), we conducted a secondary analysis of the data shared by Huensch (2022) using a model-based approach to estimate inhibition scores. We formulated three specific research questions (RQs):

RQ1: To what extent does the reliability of the three RT-based IC measures (retrieval-induced inhibition, Simon, and Stroop tasks) improve when adopting a model-based approach?

RQ2: To what extent do correlations between the three IC tasks differ when using the more reliable indices compared to traditional scoring methods?

RQ3: What are the relationships between IC measures and L2 phonological processing, based on the more reliable IC scores?

In line with principles of open science and to facilitate replication and extension of this work, all R code used for data analysis is made publicly available in the Open Science Framework (OSF; https://osf.io/bng82/).

Data set

This study utilized the data set shared by Huensch (2022), publicly available on OSF (https://osf.io/fxzvj/). The associated substantive publication is Huensch (2024). We started with the raw data sets for each of the six tasks administered (e.g., Stroop.csv)

https://doi.org/10.1017/S0272263125101095 Published online by Cambridge University Press

within the zipped folder "Data and Analysis Code.zip." In Huensch (2024), the author employed three different IC tasks, each with a unique scoring method:

1. Retrieval-induced inhibition task: Scores were calculated by dividing the median RT for inhibited items by the median RT for control items.
2. Simon task: Scores were derived by log-transforming the difference between median RTs for the congruent condition (where stimulus location matched response side) and the incongruent condition (where stimulus location conflicted with response side).
3. Stroop task: Scores were computed by log-transforming the difference between median RTs for the neutral condition (color patches) and the incongruent condition (color words printed in mismatching ink colors).

RQ1: Reliability of IC measures

Methods

Data preparation

We first applied accuracy-based screening procedures following Huensch (2024), removing all incorrectly responded trials. Each data set was then split into odd-numbered and even-numbered halves. For the Simon and Stroop Tasks, which featured randomized trial orders for each participant, we restructured the data to ensure comparability. Specifically, we reordered trials so that only those with identical content, i.e., same location and same text for the Simon task and same color and same text for the Stroop task, were considered duplet items. In the Simon task, this meant pairing trials featuring boxes of the same color in the same screen location. For the Stroop task, we paired trials presenting the same word in the same ink color.

Approach 1 (Huensch's method)

For the Retrieval-Induced Inhibition task, we calculated the RT division for each item. In the case of the Simon and Stroop tasks, we computed log-transformed RT differences. These values were then aggregated across participants, and we obtained median RT differences, conducting only by-participant analyses as per Huensch (2024). Split-half reliability was then estimated between the two halved data sets using two methods. First, we calculated Pearson correlation coefficients using the cor.test() function (*stats* package; R Core Team, 2021). Second, to account for potential outliers and abnormal distributions, we computed percentage bend correlation coefficients, which were more robust (Wilcox, 1994), using the pbcor() function (*WRS2* package; Mair & Wilcox, 2020).

Approaches 2 and 3 (Model-based methods)

For both model-based approaches, we fit linear mixed effects models for each half of the datasets. These models included trial type as a fixed effect, with item and participant as random effects, allowing for random slopes for trial type (Baayen et al. 2008). The key difference between the two approaches lies in the specification of the outcome variable. For Approach 2, we used log-transformed RT (log[RT]), while for Approach 3, we used inverse-transformed RT (–1/RT). In both cases, we maintained a maximal random-effects structure (Barr et al., 2013), using the *nloptwrap* optimizer (*optimx* package; Nash & Varadhan, 2011) and the partial Bayesian method (*blme* package; Chung et al., 2013) to address convergence issues, following Hui and Wu (2024). After fitting the models, we

https://doi.org/10.1017/S0272263125101095 Published online by Cambridge University Press

extracted by-participant random slopes and computed Pearson and robust correlations between the two halved data sets to assess reliability in the same way as in Approach 1.

We chose to explore two different data transformation methods—logarithmic and inverse—for several reasons. First, there is no universally optimal approach to data transformation given nonnormal data (see Maie et al., 2024). The choice of transformation can depend on the specific characteristics of the data and the nature of the research question. Second, logarithmic and inverse transformations are among the most commonly used methods in RT research, each with its strengths in addressing different types of distributional issues (Jiang, 2013). By including both, we can assess the robustness of our findings across different analytical approaches. Finally, comparing these two methods allows us to demonstrate how the choice of data transformation may or may not influence the results, providing insights into the methodological considerations researchers should keep in mind when analyzing RT data.

We selected the optimal transformation method based on which yielded the highest split-half reliability coefficient. While this approach helps identify the most reliable method for each task, we acknowledge the inherent subjectivity in this selection process. To address this limitation, we report all transformations tested and their resulting reliability coefficients, allowing readers to evaluate the magnitude of improvements across different approaches.

Results

Table 1 presents the correlation coefficients from the three IC tasks using the three computational approaches. A striking observation is the substantial variation in reliability estimates for the same task depending on the transformational and computational approach used. This variability is evident across all three tasks, with reliability coefficients ranging from near zero (indicating no reliability) to .73 (approaching acceptable reliability).

Consistent with previous research (e.g., Hui & Wu, 2024; Rouder & Haaf, 2019), all three IC tasks demonstrated an improvement in reliability of .20 to .50 when model-based approaches were applied. This increase essentially saved the Simon and the Stroop tasks. In other words, it brought the reliability coefficients of the Simon and Stroop tasks from unacceptably low levels (.14 to .34) to values approaching or exceeding .70, reaching what Brown (2014) considers the minimum threshold for acceptable reliability. These results also align with findings from Rouder and Haaf (2019), who reported a similar increase of approximately .20 in test-retest reliability when using model estimates compared to non-model-based methods in their reanalysis of Hedge et al.'s (2018) Stroop and Flanker task data.

Table 1. Split-half correlations for the three RT-based inhibition tasks data sets

	Approach 1 Huensch's method		Approach 2 LMM log(RT)		Approach 3 LMM (–1/RT)	
IC Task	Pearson	Robust	Pearson	Robust	Pearson	Robust
Retrieval-induced inhibition	–.23 (.37[a])	–.17 (.29[a])	.30	**.36**	.09	.06
Simon	.14	.19	.42	.45	.72	**.73**
Stroop	.34	.31	**.72**	.66	.37	.33

Note: Boldfaced values are the highest reliability coefficients observed for each task across all computational approaches.
[a]Coefficients were corrected from negative values using the method described by Krus and Helmstadter (1993) and used in Buffington et al. (2021).

https://doi.org/10.1017/S0272263125101095 Published online by Cambridge University Press

Moreover, it is important to note that the optimal data transformation differed between tasks. In the case of the Simon task, an inverse transformation performed best, while a log transformation was optimal for the Stroop task. The observed differences in optimal transformations across tasks show that researchers must be transparent about transformation selection criteria and ideally preregister their analytical decisions. Without such safeguards, there is indeed a risk of researchers trying multiple methods and selectively reporting only those yielding favorable results. In our study, we report all transformations tested to provide full transparency.

Also, despite consistent improvement across tasks, not all measures reached acceptable levels of reliability. This suggests that the model-based approach is useful to various degrees, highlighting the need for task-specific considerations in reliability analysis.

RQ2: Correlations between IC tasks

Although researchers commonly use the umbrella term "inhibitory control," this term encompasses distinct cognitive processes that likely rely on different neural mechanisms and serve different functions. The three tasks examined in this study target distinctly different aspects of IC: retrieval-induced inhibition measures unintentional resistance to proactive interference with a strong memory component; the Stroop task assesses the ability to suppress automatic responses in a linguistic context; and the Simon task evaluates domain-general spatial response inhibition. Given these substantial differences in what each task measures, strong correlations between them would not be expected even with perfect measurement. However, traditional measurement approaches may underestimate any existing relationships due to reliability issues. Our model-based approach aims to reduce measurement error to reveal the true extent of relationships (or lack thereof) between these different aspects of IC. Rather than expecting strong correlations, our goal is to determine whether more reliable measurement might reveal modest relationships that were previously obscured by measurement error, or confirm that these distinct aspects of IC function independently.

Methods

To address RQ2, we investigated the correlations among the three IC tasks using two approaches. The first approach utilized the original scores calculated following Huensch (2024), while the second employed the model-based individual random slopes identified as the most reliable from RQ1. We selected the approach that is considered methodologically preferable a priori based on established psychometric principles, i.e., that a more reliable measurement provides more accurate estimates of true relationships between constructs by minimizing the attenuating effects of measurement error. As demonstrated in our reliability analyses (see Table 1), the approach that had the highest reliability coefficients was used for each task to reveal the true between-task relationships potentially.

We used the cor.test() function for Pearson correlation coefficients and the pbcor() function from the WRS2 package for the more robust percentage bend correlation coefficients. This dual correlation analysis strategy allowed us to assess the relationships between tasks while accounting for potential outliers or nonnormality in the data.

Results

The correlation analyses revealed marked differences between the two approaches in assessing the relationships among the three IC tasks. Tables 2 and 3 present correlation

https://doi.org/10.1017/S0272263125101095 Published online by Cambridge University Press

Table 2. The correlation between tasks based on Huensch's scoring methods

	Pearson	Robust
Retrieval-Induced~ log(Simon)	r = .08 [–.18, .33], p = .54	r = .07 [–.20, .34], p = .61
Retrieval-Induced ~ log(Stroop)	r = .05 [–.21, .30], p = .71	r = .05 [–.24, .32], p = .69
log(Stroop) ~ log(Simon)	r = –.06 [–.32, .20], p = .64	r = –.09 [–.33, .15], p = .50

Table 3. The correlation between tasks based on the most reliable model-based individual random slopes

	Pearson	Robust
log(Retrieval-Induced)~ inverse(Simon)	r = .33 [.07, .54], p = .01	r = .29 [.03, .52], p = .03
log(Retrieval-Induced) ~ log(Stroop)	r = .18 [–.09, .42], p = .18	r = .24 [–.05, .47], p = .24
log(Stroop) ~ inverse(Simon)	r = .26 [0, .49], p = .05	r = .27 [.02, .47], p = .04

matrices from Huensch's original scoring methods and the most reliable model-based individual random slopes, respectively.

Using Huensch's original method, we observed consistently low correlations between tasks, with all coefficients falling below .10 (see Table 2). These results suggest little to no relationship between the three IC measures when using traditional scoring methods. Given that these tasks are designed to measure different aspects of IC, low correlations between them may align with theoretical expectations. Nevertheless, the lack of correlation raises questions about whether IC should be conceptualized as a unified IC construct with related subcomponents or as fundamentally distinct cognitive processes that share only nomenclature.

In contrast, the model-based approach yielded notably higher correlations, revealing previously undetected relationships between the tasks (see Table 3). Significant positive correlations emerged between the retrieval-induced inhibition and Simon tasks (r = .29 to .33, p = .01 to .03) and between the Simon and Stroop tasks (r = .26 to .27, p = .04 to .05). The correlation between the retrieval-induced inhibition and Stroop tasks, while higher than in the original method, did not reach statistical significance (r = .18 to .24, p = .18 to .24).

Figure 2 provides a visual comparison of the correlation coefficients and their 95% confidence intervals between three IC tasks, contrasting the results based on Huensch's (2024) scoring methods and the most reliable model-based random slopes. This visualization clearly illustrates the enhanced inter-task relationships revealed by the model-based approach.

These findings highlight the potential impact of the analytical approach on the observed relationships between IC measures. The model-based approach uncovered moderate correlations between tasks that were not apparent using traditional scoring methods, suggesting that it may provide a more sensitive measure of the shared variance between different IC tasks. This improved sensitivity could have important implications for our understanding of IC as a construct and its measurement in L2 research.

RQ3: IC predicting phonological processing

Methods

To address RQ3, we investigated whether the lack of a significant relationship between IC and L2 phonological processing reported by Huensch (2024) persisted when using

https://doi.org/10.1017/S0272263125101095 Published online by Cambridge University Press

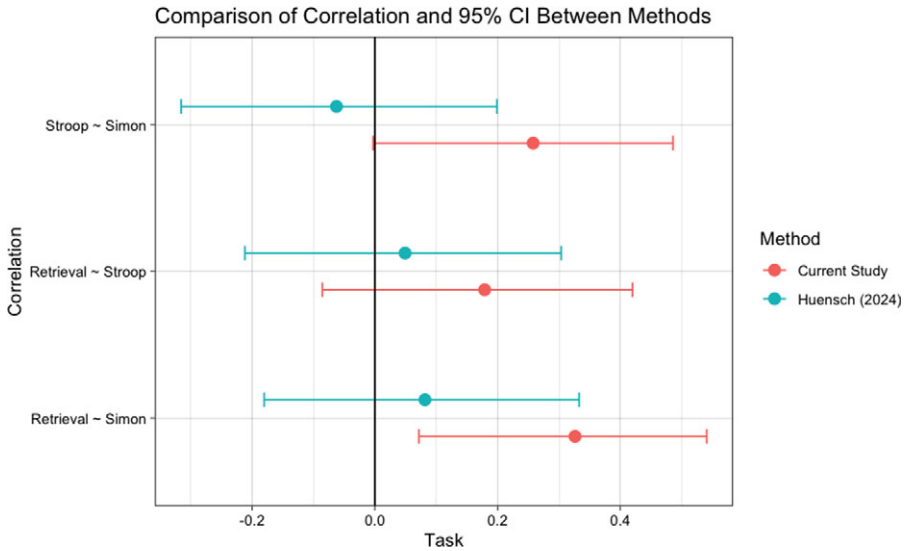

Figure 2. Comparison of correlation coefficients and 95% confidence intervals between original and model-based methods for IC tasks.
Note: This figure is based on the Pearson correlation, as the results from the robust correlation method are similar to those of the Pearson correlation method.

the most reliable IC measures derived from model-based individual random slopes. We employed two analytical approaches.

Nonparametric Spearman partial correlation analysis

We conducted analyses comparing the three IC tasks' results and the phonological scores described in Huensch (2024). For each task, we ran two analyses: one using Huensch's original RT-difference scores and another using the model-based estimates of individual random slopes that yielded the highest reliability in RQ1.

Hierarchical regression analysis

We ran two sets of analyses. First, we followed Huensch's original analysis to establish a baseline for comparison and then ran the same models but using individual random slopes generated from the most reliable model-based approach for the retrieval-induced inhibition task. In the second set of analyses, we involved both the Stroop and Simon tasks in the model selection, given that they showed the greatest improvements in reliability. In other words, we conducted comprehensive hierarchical regression analyses that included individual random slopes from these tasks to predict the vowel perception error rates. All predictors were standardized using the scale() function to aid interpretation.

The analysis began with all IC measures entered as predictors in the first step. Before model selection, we conducted model diagnostics and removed outliers (three to four) to ensure compliance with model assumptions[1]. At each later step, we identified and

[1]We conducted examinations of outliers and influential points using multiple diagnostics, including Cook's distance (with a threshold of 4/N; Altman & Krzywinski, 2016), leverage values, and standardized residuals. See the code in the OSF repository for more information.

https://doi.org/10.1017/S0272263125101095 Published online by Cambridge University Press

removed the predictor with the highest *p*-value, followed by model diagnostics, to verify that there was no violation of assumptions. We also tested if removing the predictor in model selection resulted in a significantly worse model fit via a likelihood ratio test using the anova() function in R. The best model had the optimal fit and was parsimonious. This approach allowed us to evaluate whether different IC measures would predict vowel perception differently and whether the relationship between the IC measures and vowel perception holds across the various analytical methods.

To ensure the robustness of our final models, we conducted three sensitivity analyses: two robust regression models and a bootstrap analysis. The robust regression analyses employed the rlm() function from the MASS package (Venables & Ripley, 2002) to handle potential influential observations and the lm_robust() function from the *estimatr* package to address potential heteroscedasticity. Additionally, we implemented a bootstrap analysis with 1,000 resamples using lm.boot() from the *simpleboot* package to obtain estimates that do not rely on parametric assumptions about the error distribution.

Results

Partial correlations

Table 4 compares partial correlations between IC measures and L2 phonological processing using Huensch's original scores and the most reliable model-based random slopes. Despite slight increases in some correlations using the model-based approach, all correlations remained weak (below .25), suggesting no substantial improvement in the relationship between IC and phonological processing measures.

Hierarchical regression

We first ran Huensch's (2024) models with the original retrieval-induced inhibition measure and then with the slightly improved measure using the model-based approach. The analysis with the more reliable measure showed a marginally nonsignificant effect of the retrieval-induced inhibition task, and the variance in the outcome explained by the model remained tiny (adjusted $R^2 = .07$, $\beta = -.03$, $t = -2.00$, $p = .051$).

Then, we expanded upon Huensch's original analysis (2024) first by incorporating all three IC tasks from both the original and the model-based IC scores. In both cases, the Stroop task survived model selection but was not a significant predictor. Table 5 presents the primary results of the basic models and those of the expanded analyses (see the appendix for the comparison between the standard model results and the sensitivity analyses).

Discussion

In this study, we conducted a secondary analysis of data from Huensch (2022, 2024), applying the methods tested in Hui and Wu (2024) to examine their potential to improve task reliability estimates and their implications on subsequent analyses. Our investigation yielded three key findings. First, model-based approaches could enhance reliability compared to methods using raw RT differences, particularly for the Stroop and Simon tasks and when paired with specific data transformation methods. Second, model-based random slopes for the three tasks showed numerically higher correlations between the tasks than when traditional approaches were used. Lastly, the improved task reliability only partially extended to analyses of the relationships between IC and

https://doi.org/10.1017/S0272263125101095 Published online by Cambridge University Press

Table 4. Partial correlations with Huensch's (2024) score and the most reliable model-based individual random slopes.

	Retrieval-induced		Simon		Stroop	
	Huensch	log(RT)	Huensch	–1/(RT)	Huensch	log(RT)
ABX error (vowels)	*r* = –.13 [–.41, .16], *p* = .38	*r* = –.23 [–.48, .05], *p* = .11	*r* = –.12 [–.37, .16], *p* =.41	*r* = –.23 [–.47, .05], *p* = .11	*r* = .18 [–.08, .42], *p* =. 17	*r* = –.20 [–.44, .07], *p* = .15
ABX error (consonants)	*r* = –.11 [–.37, .17], *p* = .45	*r* = –.01, [–.30, .28], *p* = .96	*r* = .07 [–.17, .30], *p* = . 59	*r* = –.15 [–.40, .11], *p* = .26	*r* = .23 [–.02, .45], *p* =. 07	*r* = –.25 [–.49, .03], *p* = .08
Vowel production *z* score	*r* = .09, [–.19, .35], *p* = .53	*r* = .08 [–.19, .34], *p* = .55	*r* = –.18 [–.44, .11], *p* =. 22	*r* = –.09 [–.35, .18], *p* = .50	*r* = .13 [.13, .38], *p* =. 33	*r* = –.02 [–.27, .23], *p* = .86
Consonant production	*r* = .07 [–.20, .33], *p* = .61	*r* = .16 [–.12, .42], *p* = .26	*r* = .01 [–.24, .27], *p* =. 92	*r* = .12 [–.16, .38], *p* = .40	*r* = .10 [–.16, .34], *p* =. 47	*r* = –.004 [–.28, .27], *p* = .98

Table 5. Hierarchical regression results with retrieval-induced inhibition predicting vowel perception error rates while controlling for proficiency with different methods

Analysis	Step	Predictor	R^2	β	95% CI	t	p
Huensch's original IC measure *with only 1 IC task*	Step 1	X_Lex	.03	–.03	–.06, .01	–1.39	.17
		Inhibition		–.02	–.06, .01	–1.16	.25
	Step 2	X_Lex	.02	–.03	–.06, .01	–1.54	.13
Huensch's original IC measure *model-based and with only 1 IC task*	Step 1	X_Lex	.07	–.03	–.06, .01	–1.60	.12
		Inhibition		–.03	–.07, .00	–2.00	.05
	Step 2	Inhibition	.05	–.03	–.07, .00	–1.95	.06
Huensch's original IC measure *with all 3 IC tasks*	Step 1	X_Lex	.12	–.03	–.06, .00	–1.92	.06
		Inhibition		–.02	–.05, .01	–1.44	.16
		log(Simon)		–.00	–.03, .03	–.00	1.00
		log(Stroop)		.02	–.00, .05	1.75	.09
	Step 2	X_Lex	.14	–.03	–.06, .00	–1.94	.06
		Inhibition		–.02	–.05, .01	–1.47	.15
		log(Stroop)		.02	–.00, .05	1.78	.08
	Step 3 - *best model*	X_Lex	.12	–.03	–.06, –.01	–2.32	.02*
		log(Stroop)		.02	–.00, .05	1.71	.09
	Step 4	X_Lex	.09	–.04	–.07, –.01	–2.49	.02*
Huensch *model-based and with all 3 IC tasks*	Step 1	X_Lex	.17	–.04	–.07, –.01	–2.79	.01**
		Inhibition		–.02	–.05, .02	–.96	.34
		Simon		–.02	–.05, .02	–1.03	.31
		Stroop		–.02	–.05, .01	–1.39	.17
	Step 2	X_Lex	.17	–.04	–.07, .01	–2.83	.01**
		Simon		–.02	–.05, .01	–1.30	.20
		Stroop		–.02	–.05, .01	–1.57	.12
	Step 3 - *best model*	X_Lex	.16	–.04	–.07, –.01	–2.83	.01**
		Stroop		–.03	–.06, .00	–1.92	.06
	Step 4	X_Lex	.11	–.04	–.07, –.01	–2.74	.01**

phonological processing. Compared to the original methods in Huensch (2024), the application of the model-based approach on neither correlation analyses nor hierarchical analyses showed any meaningful differences, indicating that the null effects reported in Huensch (2024) were not likely due to the unreliability of the IC tasks.

Addressing RQ1, we found that model-based approaches generally improved reliability, particularly for the Simon and Stroop tasks. While these tasks showed low reliability using the standard approaches (confirming Huensch, 2024 and Hedge et al., 2018, 2022), the model-based approach increased reliability by .20 to .50 across tasks, with both the Stroop and Simon tasks reaching near-acceptable levels (.72–.73). These results align with the findings reported in Rouder and Haaf (2019) (i.e., from .55 to .72 for the full set of Stroop tasks), showing that model-based approaches could improve the reliability of IC tasks when they are used as predictors of L2 learning outcomes.

However, the retrieval-induced inhibition task showed limited improvement, remaining at a low level of reliability (.30–.36). This persistent issue suggests that some measures may be inherently more susceptible to measurement error and reduced between-participant reliability, regardless of the statistical approach used. As Hui and Wu (2024) noted, not all datasets benefit equally from model-based approaches, with item variability, for example, potentially moderating their effectiveness. These different results highlight the importance of careful selection of data analysis procedures in individual differences research in SLA and beyond. As demonstrated in our analysis and supported by Hui and Wu (2024), model-based approaches yield greater benefits when applied to tasks with substantial item variability. For tasks like the Stroop and Simon that involve stimuli with varying difficulty levels, model-based approaches

https://doi.org/10.1017/S0272263125101095 Published online by Cambridge University Press

may be more easily ready to partition this variance. To determine whether the model-based approach could be helpful, researchers can (1) consider task structure and item characteristics, (2) evaluate the sample size and trial numbers, and (3) conduct pilot reliability assessments before a full-scale implementation. Moreover, a potential explanation for the retrieval-induced inhibition task's resistance to improvement may lie in its unique calculation method: unlike the Stroop and Simon tasks that use RT difference (e.g., in Stroop, IC score = average RT – neutral RT), this task expresses IC in terms of a ratio (IC score = inhibited RT/control RT). This computational difference calls for further methodological work that investigates the optimal approaches to the computation of an IC score for different tasks and its implications for subsequent analyses to address, for example, the predictive validity of the measure.

Notably, the effectiveness of the model-based approach was somewhat moderated by the data transformation method employed in the sense that the optimal transformation (log *vs.* inverse) varied by task. Log transformation was proven most useful for the Stroop task, while the Simon task benefited most from inverse transformation. This finding raises the question of how to determine the "best" analytical approach when multiple transformations are available. In our study, we used reliability coefficients as the primary selection criterion, with higher values indicating better measurement precision. However, this approach requires careful consideration to avoid the potential pitfalls of researchers' degrees of freedom.

In SLA, response and processing times are often transformed in statistical analyses, but the choice of which transformation to use is not always justified. Few studies conduct sensitivity analysis involving more than one transformation method. Notable exceptions include Maie et al. (2024), who demonstrated the impact of arbitrary choices in analysis (e.g., transformation) on the results of a study, and Wu and Toda Cosi (2025), who showed that certain cases do not benefit from standard transformation methods and require alternative modeling approaches. In the present context of reliability estimation, we encourage researchers to consider the resulting levels of reliability, in addition to model diagnostics, in making transformation decisions. That said, researchers must be transparent in their election and not abuse their researcher's degree of freedom to arrive at desirable results by: (1) preregistering their analytical plans including transformation methods before data collection; (2) reporting all transformations tested rather than only the "optimal" one; (3) establishing clear criteria for what constitutes meaningful reliability improvement before analysis begins. Also, more methodological investigations should be carried out to help applied researchers make more informed decisions when selecting an appropriate approach. We suggest the following steps.

1. Begin with theoretical considerations about the distribution of your RT data. Log transformations are typically more appropriate for positively skewed distributions (Feng et al., 2014), while inverse transformations may better handle extreme outliers (Özdemir & Çavuş, 2016).
2. Apply prescreening criteria before selecting transformations. For example, evaluate whether transformations effectively normalize residuals and check variance homogeneity using diagnostic plots.
3. Report reliability estimates for all transformations tested, not just the optimal one.
4. Conduct sensitivity analyses to determine whether your conclusion remains stable across different transformation approaches.

Although the model-based approach improved the reliability of individual measures (as shown in RQ1), it revealed only slightly stronger relationships between the tasks. Using Huensch's original method, correlations between tasks were consistently low

https://doi.org/10.1017/S0272263125101095 Published online by Cambridge University Press

(below .10). In contrast, the model-based approach yielded notably higher correlations, with significant positive correlations observed between the retrieval-induced inhibition and Simon tasks (r = .29 to .33) and between the Simon and Stroop tasks (r = .26 to .27). This suggests that the three tasks may be tapping into somewhat different aspects of IC. This result has theoretical implications for the construct of IC, because it has sometimes been, at least implicitly, conceptualized as single-dimensional. Indeed, when justifying her addition of the Stroop and Simon tasks, Huensch (2024) argued that the three different tasks measure distinct aspects of IC: the retrieval-induced inhibition task taps into unintentional resistance to proactive interference with a strong language processing component; the Stroop task examines language-oriented but not language-focused response inhibition; and the Simon task assesses domain-general inhibition. The low inter-task correlations require researchers using IC measures to be very specific about their target subconstruct of IC, because the findings can depend on task selection. More generally, this specificity is important because one goal of Instructed SLA is to examine how specific individual differences interact with treatment to influence language learning (DeKeyser, 2021).

Despite the improved reliability of IC measures, the relationship between IC and L2 phonological processing remained weak to moderate, according to the partial correlation analyses and the hierarchical regression analyses with multiple sensitivity checks (RQ3). This finding further confirmed the null effects reported in Huensch (2024). In addition, although our inclusion of the Stroop and Simon tasks has revealed some role of the Stroop task in accounting for vowel perception, the effects were not significant, and the variance explained was almost negligible. At the surface level, our model-based approach did not change the conclusion drawn by Huensch (2024), leading some to wonder about its usefulness. At the same time, the key contribution of the secondary analysis is that, again, with the more reliable measures derived from a mixed-effects model, researchers can rule out the possibility of the null effects resulting from low reliability. In other words, these findings are methodologically important, as they demonstrate that improved reliability of predictor measures, through the use of a model-based approach, has the potential to address confounding statistical issues (Hedge et al., 2018; Rey-Mermet et al., 2018) and provide a clearer assessment of these measures' true predictive power or lack thereof.

These findings have significant implications for the broader field of SLA research beyond IC studies. First, they highlight the importance of measurement reliability when investigating individual differences in cognitive abilities that may influence language acquisition. Researchers exploring cognitive predictors of learning outcomes should prioritize reliable measurement to better uncover genuine relationships. Second, our work demonstrates how advanced statistical methods can be productively applied to existing datasets in SLA, allowing researchers to extract additional insights from previously collected data—a practice that aligns with growing emphasis on open science and resource efficiency in our field. Finally, the methodological advances demonstrated here extend beyond IC to any SLA research involving RT-difference measures, including studies of lexical access, syntactic processing, and language comprehension, offering new analytical tools to enhance the rigor of future investigations across diverse domains of SLA research.

Conclusion

Throughout the article, we have repeatedly underscored the importance of considering measurement reliability when studying individual differences in cognitive abilities and their relationships to language learning. As suggested by previous research (Buffington

https://doi.org/10.1017/S0272263125101095 Published online by Cambridge University Press

et al., 2021; Hui & Wu, 2024), the low reliability of a task can be partially attributed to computational methods, and more robust approaches, such as the model-based method employed in this study, may provide an alternative for L2 researchers. It is important to keep in mind that reliability is a prerequisite of validity (Davis, 1992; McKay & Plonsky, 2021) and represents a cornerstone in quantitative research. Since RT-difference measures can be very unreliable, as many external factors can influence the data (e.g., handedness, physical difficulties, and coordination, Hui & Jia, 2024), and much SLA research relies on RT data, serious researchers should not turn a blind eye to the issues surrounding reliability.

Competing interests. We have no known conflict of interest to disclose.

References

Altman, N., & Krzywinski, M. (2016). Analyzing outliers: Influential or nuisance? *Nature Methods*, *13*(4), 281–282. https://doi.org/10.1038/nmeth.3812

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Buffington, J., Demos, A. P., & Morgan-Short, K. (2021). The reliability and validity of procedural memory assessments used in second language acquisition research. *Studies in Second Language Acquisition*, *43*(3), 635–662. https://doi.org/10.1017/S0272263121000127

Chung, Y., Rabe-Hesketh, S., Dorie, V., Gelman, A., & Liu, J. (2013). A nondegenerate penalized likelihood estimator for variance parameters in multilevel models. *Psychometrika*, *78*(4), 685–709. https://doi.org/10.1007/s11336-013-9328-2

Cohen, L., Manion, L., & Morrison, K. (2017). Validity and reliability. In L. Cohen, L. Manion, & K. Morrison, *Research Methods in Education* (8th ed., pp. 245–284). Routledge. https://doi.org/10.4324/9781315456539-14

Darcy, I., Mora, J. C., & Daidone, D. (2016). The role of inhibitory control in second language phonological processing: Inhibitory control and L2 phonology. *Language Learning*, *66*(4), 741–773. https://doi.org/10.1111/lang.12161

Davis, K. A. (1992). Validity and reliability in qualitative research on second language acquisition and teaching: Another researcher comments. *TESOL Quarterly*, *26*, 605–608. https://doi.org/10.2307/3587190

DeKeyser, R. M. (2021). Aptitude-treatment interaction in second language learning: Introduction to the special issue. In R. M. DeKeyser (Ed.), *Benjamins Current Topics* (Vol. *116*, pp. 1–4). John Benjamins Publishing Company. https://doi.org/10.1075/bct.116.01dek

Fang, S., & Wu, Z. (2022). Syntactic prediction in L2 learners: evidence from English disjunction processing. *International Review of Applied Linguistics in Language Teaching*, *62*(2), 429-456. https://doi.org/10.1515/iral-2021-0223

Feng, C., Wang, H., Lu, N., Chen, T., He, H., Lu, Y., & Tu, X. M. (2014). Log-transformation and its implications for data analysis. *Shanghai archives of psychiatry*, *26*(2), 105–109. https://doi.org/10.3969/j.issn.1002-0829.2014.02.009

Frinsel, F. F., & Christiansen, M. H. (2024). Capturing individual differences in sentence processing: How reliable is the self-paced reading task? *Behavior Research Methods*, *56*(6), 6248–6257. https://doi.org/10.3758/s13428-024-02355-x

Hedge, C., Powell, G., Bompas, A., & Sumner, P. (2022). Strategy and processing speed eclipse individual differences in control ability in conflict tasks. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *48*(10), 1448–1469. https://doi.org/10.1037/xlm0001028

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods*, *50*(3), 1166–1186. https://doi.org/10.3758/s13428-017-0935-1

Hedge, C., Vivian-Griffiths, S., Powell, G., Bompas, A., & Sumner, P. (2019). Slow and steady? Strategic adjustments in response caution are moderately reliable and correlate across tasks. *Consciousness and Cognition*, *75*, 102797. https://doi.org/10.1016/j.concog.2019.102797

Huensch, A. (2022). *Clarifying the role of inhibition in L2 phonological processing: A close replication of Darcy et al. (*2016*)* [Data set]. Open Science Framework. https://doi.org/10.17605/OSF.IO/FXZVJ

Huensch, A. (2024). Clarifying the role of inhibitory control in L2 phonological processing: A preregistered, close replication of Darcy et al. (2016). *Studies in Second Language Acquisition*, *46*(5), 1392–1412. https://doi.org/10.1017/S0272263124000238

Hui, B., Godfroid, A., & Elgort, I. (2025). *A construct validation study of time-sensitive word knowledge measures. Applied Linguistics*. Advance online publication. https://doi.org/10.1093/applin/amaf037; https://doi.org/10.31219/osf.io/dwjmn

Hui, B., & Jia, R. (2024). Reflecting on the use of response times to index linguistic knowledge in SLA. *Annual Review of Applied Linguistics*, Advance online publication. doi:10.1017/S0267190524000047

Hui, B., & Wu, Z. (2024). Estimating reliability for response-time difference measures: Toward a standardized, model-based approach. *Studies in Second Language Acquisition*, *46*(1), 227–250. https://doi.org/10.1017/S027226312300027X

Jiang, N. (2013). *Conducting reaction time research in second language studies*. Routledge.

Krus, D. J., & Helmstadter, G. C. (1993). The Problem of Negative Reliabilities. *Educational and Psychological Measurement*, *53*(3), 643-650. https://doi.org/10.1177/0013164493053003005

Maie, R. (2022). *Testing the three-stage model of second language skill acquisition* (Publication number 2748315725) [Doctoral Dissertation, Michigan State University]. ProQuest Dissertations & Theses Global. https://www.proquest.com/dissertations-theses/testing-three-stage-model-second-language-skill/docview/2748315725/se-2

Maie, R., Eguchi, M., & Uchihara, T. (2024). Arbitrary choices, arbitrary results: Three cases of multiverse analysis in L2 research. *Research Methods in Applied Linguistics*, *3*(2), 100124. https://doi.org/10.1016/j.rmal.2024.100124

Mair, P., & Wilcox, R. (2020). Robust statistical methods in R using the WRS2 package. *Behavior Research Methods*, *52*(2), 464–488. https://doi.org/10.3758/s13428-019-01246-w

McKay, T., & Plonsky, L. (2021). Reliability analyses: Estimating error in L2 research. In P. M. Winke & T. Brunfaut (Eds.), *The Routledge handbook of second language acquisition and language testing* (pp. 468–482). Routledge Taylor and Francis Group.

Nash, J. C., & Varadhan, R. (2011). Unifying optimization algorithms to aid software system users: Optimx for *R*. *Journal of Statistical Software*, *43*(9). https://doi.org/10.18637/jss.v043.i09

Özdemir, Ö., & Çavuş, M. (2016). Performance of the inverse transformation method for extreme value distributions. In *X^th International Statistics Days Conference (ISDC'2016), Giresun, Turkey* (Vol. 8).

Porte, G., & McManus, K. (2019). *Doing replication research in applied linguistics* (1st ed.). Routledge. https://doi.org/10.4324/9781315621395

R Core Team. (2021). *R: A Language and Environment for Statistical Computing* [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Rey-Mermet, A., Gade, M., & Oberauer, K. (2018). Should we stop thinking about inhibition? Searching for individual and age differences in inhibition ability. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(4), 501–526. https://doi.org/10.1037/xlm0000450

Rouder, J. N., & Haaf, J. M. (2019). A psychometrics of individual differences in experimental tasks. *Psychonomic Bulletin & Review*, *26*(2), 452–467. https://doi.org/10.3758/s13423-018-1558-y

Tan, L. C., & Yap, M. J. (2016). Are individual differences in masked repetition and semantic priming reliable? *Visual Cognition*, *24*(2), 182–200. https://doi-org.proxy-um.researchport.umd.edu/10.1080/13506285.2016.1214201

Unsworth, N., & Engle, R. W. (2005). Individual differences in working memory capacity and learning: evidence from the serial reaction time task. *Memory & Cognition*, *33*(2), 213–220. https://doi.org/10.3758/bf03195310

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed.). Springer.

Wilcox, R. R. (1994). The percentage bend correlation coefficient. *Psychometrika*, *59*(4), 601–616. https://doi.org/10.1007/BF02294395

Wu, Z., & Toda Cosi, M. (2025). From lab to web: Replicating cross-language translation priming asymmetry in an online environment. *Research Methods in Applied Linguistics*, *4*(3), 100247. https://doi.org/10.1016/j.rmal.2025.100247

Zhang, N., Wu, Z., & Wang, M. (2025). Cross-language phonological activation in bilingual visual word recognition: A meta-analysis. *Psychonomic Bulletin & Review*. https://doi.org/10.3758/s13423-025-02692-8

Appendix Sensitivity check for RQ3

Comparison of hierarchical regression results from different sensitivity analysis methods with retrieval-induced inhibition predicting vowel perception error rates while controlling for proficiency with different methods.

	Step	Predictor	R^2	β	95% CI	t	p
Huensch's original IC measure	***best model***	X_Lex	.12	–.03	–.06, –.01	–2.32	.02*
		log(Stroop)		.02	–.00, .05	1.71	.09
	Sensitivity check - *rlm robust regression*	X_Lex	NA	–.04	–.07, –.00	–2.22	.03*
		log(Stroop)		.03	–.00, .06	1.82	.07
	Sensitivity check - *lm_robust regression*	X_Lex	.17	–.04	–.07, –.01	–2.48	.02*
		log(Stroop)		.03	.00, .06	2.27	.03*
Model-based IC measure	***best model***	X_Lex	.16	–.04	–.07, –.01	–2.83	.01**
		Stroop		–.03	–.06, .00	–1.92	.06
	Sensitivity check - *rlm robust regression*	X_Lex	NA	–.04	–.07, –.01	–2.52	.02*
		Stroop		–.03	–.06, .00	1.74	.09
	Sensitivity check - *lm_robust regression*	X_Lex	.16	–.04	–.07, –.01	–2.77	.01**
		Stroop		.03	–.06, .00	–1.92	.06

Cite this article: Wu, Z., Jia, R., & Hui, B. (2025). Saving the reliability of inhibitory control measures? An extension of Huensch (2024) and Hui and Wu (2024). *Studies in Second Language Acquisition,* 1–21. https://doi.org/10.1017/S0272263125101095