Studies in Second Language Acquisition (2023), 1–24
doi:10.1017/S027226312300027X

METHODS FORUM

Estimating reliability for response-time difference measures: Toward a standardized, model-based approach

Bronson Hui and Zhiyi Wu

Graduate Program of Second Language Acquisition, School of Languages, Literatures, and Cultures, University of Maryland, College Park, MD, USA
Corresponding author: Bronson Hui; Email: bhui@umd.edu

(Received 11 January 2023; Revised 17 April 2023; Accepted 24 April 2023)

Abstract

A slowdown or a speedup in response times across experimental conditions can be taken as evidence of online deployment of knowledge. However, response-time difference measures are rarely evaluated on their reliability, and there is no standard practice to estimate it. In this article, we used three open data sets to explore an approach to reliability that is based on mixed-effects modeling and to examine model criticism as an outlier treatment strategy. The results suggest that the model-based approach can be superior but show no clear advantage of model criticism. We followed up these results with a simulation study to identify the specific conditions in which the model-based approach has the most benefits. Researchers who cannot include a large number of items and have a moderate level of noise in their data may find this approach particularly useful. We concluded by calling for more awareness and research on the psychometric properties of measures in the field.

Introduction

Applied second language (L2) researchers have been using tasks based on response time (RT) to tap into learners' grammar and vocabulary knowledge (e.g., Elgort, 2011; Granena, 2013) as well as individual differences in attributes such as procedural memory capacity (e.g., Buffington et al., 2021). Some of these measures, traditionally used to index online language processing in the psycholinguistics literature, have become commonplace in second language acquisition (SLA). This is in part because they are believed to tap into the knowledge that is available for automatic processing, a fundamental basis for authentic language use (e.g., Elgort, 2011; Suzuki, 2017). Examples of such measures include word-monitoring tasks (Godfroid & Kim, 2021; Granena, 2013; Suzuki et al., 2022), self-paced reading (SPR) tasks (Fang & Wu, 2022b; Godfroid & Kim, 2021; Marsden et al., 2018), and judgment tasks in the priming paradigm (Elgort, 2011; Hui et al., 2022a; Plonsky et al., 2020). In data analysis, researchers using these measures typically focus on differences in RTs across experimental conditions. In

a word-monitoring or SPR task, for example, a slowdown in response (or processing) when encountering ungrammaticality is taken as evidence of the learner's sensitivity to anomalies, which in turn is interpreted as a manifestation of the learner deploying the relevant grammatical knowledge online.

Although these tasks have been useful, applied researchers should exercise caution. An important reason is that these tasks are often used in the psycholinguistics literature to demonstrate a group-level effect, so the extent to which these tasks function well as an individual-difference measure in applied contexts remains an open question (Draheim et al., 2019). In this light, researchers must pay more attention to the fundamental psychometric properties of these tasks, such as reliability. However, the reliability of these measures is rarely reported in both SLA and neighboring fields such as cognitive psychology (e.g., Marsden et al., 2018; Parsons et al., 2019; Plonsky & Derrick, 2016). Even when it is reported, the estimation method is not always detailed. As a result, how reliability has been and should be computed is a mystery, leaving applied researchers in a quandary because there is no obvious standard practice to follow. Such a lack of reference in the literature can severely limit researchers' ability to make strong claims about what these tasks measure and the relationship between the object of measurement and other variables of interest (McKay & Plonsky, 2021). In this article, we take a step toward an informed, standardized approach to computing reliability for RT differences by evaluating three estimation methods first with three open data sets and then with simulated data sets. A secondary goal of this paper is also to improve researchers' awareness of the importance of appropriately estimating the error associated with their instruments. Here, we present two studies, the first of which concerns a model-based approach to reliability for RT-difference measures. We then report a series of simulations that were based on the results of the first study to elucidate the specific conditions under which the model-based approach may have the most benefits.

Use of response-time difference measures

The ways in which language processing is investigated in psycholinguistics have inspired L2 researchers to apply online methods to address fundamental questions in SLA. For example, some SLA researchers have been interested in using these methods to measure implicit and automatized explicit knowledge (e.g., Bowles, 2011; Ellis, 2005; Vafaee et al., 2017) because learners can access this knowledge rapidly and effortlessly in real-time processing and fluent language use (Ellis, 2005). In this light, a direct way to tap into this kind of knowledge is to emphasize the processing of the learner in the measurement of linguistic knowledge. This emphasis has led to the rapid adoption of online measures in SLA research (e.g., Elgort et al., 2018; Godfroid, 2020; Marsden et al., 2018), especially when these tasks are now available from various data collection platforms (Patterson & Nicklin, 2023). Given the focus of the present paper on reliability, we limit our discussion to measures that involve a comparison of performance (typically response or processing times) between different experimental conditions.

First, Godfroid and Kim (2021) used both word monitoring and SPR tasks to measure implicit grammatical knowledge of six target structures (three morphological [e.g., third person *-s*] and three syntactic [e.g., embedded questions]). In the SPR task, for example, 131 English learners read grammatical or ungrammatical versions of stimulus sentences (e.g., *The old woman* ***enjoys*** *reading many different famous novels* (p. 615)). Participants were told to press a button that recorded the time that elapsed

https://doi.org/10.1017/S027226312300027X Published online by Cambridge University Press

from the previous hit to proceed to the next word. Given that the spillover region (e.g., *reading*) followed the critical grammatical feature in question, participants were expected to show sensitivity to grammatical violations if they had implicit knowledge of the target structure. This sensitivity was operationalized and measured as a slowdown in processing when encountering ungrammaticality, compared with the grammatical baseline condition. In analytical terms, the difference in RTs between the grammatical and the ungrammatical trials represented an indication of implicit knowledge (e.g., Granena, 2013; Maie & DeKeyser, 2020; Suzuki, 2017; Suzuki et al., 2022).

In some cases, researchers expect to observe a speedup in the processing of SPR times. This approach can be based on consistent evidence for faster processing of formulaic language than matched control phrases (see Siyanova-Chanturia, 2013, for an overview). This processing advantage is attributed to the entrenchment of these units in memory as a result of frequent exposures (Siyanova-Chanturia & Martinez, 2014). Therefore, faster reading times can be expected when learners read a sentence containing formulaic language than a matched control version. For example, Fang and Wu (2022b) investigated Chinese learners' knowledge of the *either–or* construction in English. In an SPR task, participants read stimulus sentences such as *Jay painted either | the big house | or the old car | for his family over the summer* (p. 9), with or without *either*. The authors reported a speedup in reading times for the critical region (in this example, *or the old car*) in the trials that contained *either*, suggesting that these learners used their knowledge of formulaic construction to predict upcoming information in reading.

In addition to SPR, judgment tasks in the priming paradigm have also been used to index learners' lexical knowledge (e.g., Elgort, 2011; Hui et al., 2022a). Priming often involves the presentation of a prime before a target. The prime is meant to influence the processing of subsequent linguistic information contained in the target due to prime–target orthographic, phonological, and/or semantic relationships (see Trofimovich & McDonough, 2011, for an overview). Elgort (2011) introduced this technique to the field as a measure of the representational aspects of vocabulary knowledge as a result of intervention in an applied context. In her study, participants learned pseudowords (e.g., *obsolate*) using flashcards. With three lexical decision tasks using different kinds of priming—namely, form priming, masked-repetition priming, and semantic priming—the author showed that deliberate learning of words can result in knowledge indexed by these implicit tasks. For example, in masked-repetition priming, participants made a faster lexical decision when the target was preceded by an identical prime (e.g., *obsolate-OBSOLATE*) than by an unrelated prime (e.g., *mythical-OBSOLATE*). Again, a faster RT in related trials, compared with unrelated trials, constituted a measure of "the formal-lexical representations of the stimuli" (p. 382). The formation of these representations was interpreted as learning products in the wake of deliberate word-learning activities.

So far, we have focused our discussion on indexing grammatical and lexical knowledge with RT difference tasks. In addition, these tasks can be used to measure attributes of individual differences such as procedural memory capacity (Buffington et al., 2021; Maie, 2022). In an alternating serial RT task, for example, participants in the study by Buffington et al. (2021) saw a picture of a dog head filling a position in a row of four circles. The participants were instructed to press the button that corresponded to the position. In all experimental trials, the target location was dictated by a predetermined sequence unknown to the participant, alternated with random positions. Specifically, the odd-numbered trials were sequenced according to a pattern, whereas the even-numbered trials were random. The idea of the design was that if participants

https://doi.org/10.1017/S027226312300027X Published online by Cambridge University Press

could learn the predetermined sequence through exposures, they were expected to respond faster in patterned, sequenced trials than in random trials. Thus, the ability to learn a hidden sequence, an attribute linked to procedural memory capacity, is quantified by a comparison of RTs in patterned versus random trials. This RT difference has been used in the literature as an individual-difference measure of procedural memory capacity (e.g., Brill-Schuetz & Morgan-Short, 2014; Faretta-Stutenberg & Morgan-Short, 2018; Maie, 2022).

Taken together, measures that involve differences in RTs have been adopted to index grammar and vocabulary knowledge, as well as cognitive attributes such as procedural memory capacity. Given the ubiquity of RT-difference measures in SLA, there is a need to consistently evaluate the measures that researchers begin to rely on, especially when they are applied to a novel context of investigation (e.g., adopted from psycholinguistics to SLA). A central aspect of such an evaluation is instrument reliability.

Instrument reliability in SLA

A cornerstone of quantitative research is measurement, which can be defined as the principled assignment of numerical values to objects, attributes, or events (e.g., Stevens, 1946). In SLA, linguistic knowledge in the L2 is perhaps the most important attribute to measure. The ways in which we measure knowledge such that, for example, a learner scoring higher on a grammar test possesses more or better grammatical knowledge than their peers is thus fundamental to our work. At the same time, measurement error is inevitable because researchers are often unable to tap into the constructs of interest directly (McKay & Plonsky, 2021). On this account, appropriately estimating the amount of error is critical because it allows researchers to understand the limitations of their instruments and represents one of the very first and most critical steps in evaluating a measure. Reliability, or consistency across measurements, has traditionally been regarded as a necessary condition for validity. As Davis (1992) put it more than three decades ago, "an unreliable measure cannot be valid" (p. 606). In other words, interpretations of scores largely assume that the individual demonstrates at least some consistency in their scores across independent measurements (American Educational Research Association et al., 2014). Indeed, McKay and Plonsky (2021) argue that claims cannot be made about what is being measured and its relationship with other variables without sufficient evidence that the score in question is consistent at acceptable levels. Therefore, reliability deserves more attention in any scientific pursuit. Otherwise, researchers could be drawing conclusions without confirming their measurement is reliable, which renders the claims potentially more questionable than they should be.

From a statistical point of view, unreliability should also be addressed to the extent that is possible. When researchers build a general linear model, predictor variables are assumed to be measured without error. Using an unreliable measure to predict an outcome then violates this assumption. Even when the RT-difference measure is the outcome of the model, the variance in the outcome that is not explained by the model can be due to (1) its own unreliability, (2) the lack of sufficient explanatory power of the predictors, or (3) a mix of both. When an unreliable measure is used as either a predictor or the outcome, researchers are losing statistical power because the true relationship between the predictors and the outcome can be masked by error. This loss of power can be critical, especially when researchers in some subfields are already struggling to have sufficiently large sample sizes (Loewen & Hui, 2021; Vitta et al., 2021). However, if researchers are able to maximize instrument reliability, we can then focus on a more substantive search for factors that can explain a phenomenon

https://doi.org/10.1017/S027226312300027X Published online by Cambridge University Press

examined by the researchers. In summary, instrument reliability plays an important role in quantitative research in SLA and should be considered in the constant evaluation of instruments.

Reliability for response-time differences

Despite the fundamental role of instrument reliability, L2 researchers often do not report the internal consistency of their tests (McKay & Plonsky, 2021; Plonsky & Derrick, 2016). Researchers using RT-difference measures, such as SPR and judgment tasks, are no exception (Marsden et al., 2018; Plonsky et al., 2020). Although non-reporting does not necessarily imply low reliability (Plonsky & Derrick, 2016), there is a growing literature that should concern researchers, particularly those relying on RT differences.

First, in psycholinguistics, Tan and Yap (2016) reported shockingly low levels of reliability in masked-repetition and semantic priming. In their study, 240 native English speakers performed tasks within the masked-repetition and semantic priming paradigms in separate experimental blocks. The authors evaluated the consistency of the measurements with two reliability measures: split-half reliability and test-retest reliability (see McKay & Plonsky, 2021, for an overview of reliability measures). The former typically involves dividing data collected within one single experimental session into two halves (e.g., odd- and even-numbered items) before computing a correlation between the performance in the two subsets of data. The latter uses data gathered from the same participant but in separate test sessions. These two reliability estimates may seem similar at first sight, but discrepancies between the two have recently been reported, suggesting that the underlying factors influencing the level of reliability estimated by these approaches might differ (Oliveira et al., 2022; West et al., 2018). In Tan and Yap (2016), the correlation coefficients for the repetition priming ranged from .21 (Pearson, test-retest) to .43 (robust, split-half). For the masked semantic priming, these figures were between .05 (Pearson, split-half) and .17 (robust, split-half). The authors cautioned that "the unreliability … makes [the measures] a poor candidate for studying individual differences" (p. 195).

In SLA, Buffington et al. (2021) reported a split-half reliability of .42 for the alternating serial RT task discussed above, based on the performance of 119 participants. Relatedly, authors who rely on serial RT tasks to investigate statistical learning (outside of SLA) have also expressed serious concerns about the poor psychometric properties of the available individual-difference measures for statistical learning (Arnon, 2019; Lammertink et al., 2020; Oliveira et al., 2022; Siegelman et al., 2017; West et al., 2017). For example, Arnon (2019) examined three statistical learning tasks (two auditory and one visual) which were administered to both children and adults twice, with a two-month gap between administrations. The test-retest reliability estimates varied from .45 to .70, depending on the task in the adult data. When correlating the different tasks, which are meant to index a similar construct (i.e., statistical learning ability), the highest figure was .41. For the child data, the picture was even more gloomy, as the test-retest reliability for individual tasks ranged from .01 to .33, with the highest correlation between tasks at .33.

In cognitive psychology more generally, tasks believed to tap into cognitive abilities, such as executive functioning, also showed only moderate test-retest reliability, as reported in Hedge et al. (2018). For example, intraclass correlations were at .40 and .57 for the RT results of a flanker task, where participants indicated the direction of an arrow in the middle of others that point in the same (congruent) or different

https://doi.org/10.1017/S027226312300027X Published online by Cambridge University Press

(incongruent) directions. Given this level of reliability, it might not be surprising that tasks that, again, are supposed to tap into the same construct (i.e., the flanker and Stroop tasks) correlated with each other at only .14 (Hedge et al., 2018). Other studies have also warned researchers about this reliability issue with RT difference tasks in individual differences research in cognition (e.g., Paap & Sawi, 2016; Verhaeghen & De Meersman, 1998).

What merits special attention here is that many of these tasks have consistently been reported to elicit robust effects at the group level but the very same task is very unreliable when used to examine individual differences. This reliability paradox (Hedge et al., 2018) urges researchers to pause and evaluate RT-difference measures. Indeed, a reviewer suggested that "the field of individual differences in cognition is experiencing somewhat of a measurement crisis." Although the extent to which this statement is true remains an open question, a serious yet simple question must be asked: What are we measuring with these tasks after all?

Potential sources of unreliability

Given the general unreliability, researchers should seek to understand the sources of unreliability. First and foremost, the unreliability can be due to substantive factors that can be related to the particular processes that the tasks seek to examine. For example, Tan and Yap (2016) argued that the psycholinguistic mechanisms underlying semantic priming are controlled, as opposed to automatic, in nature. Therefore, the performance of the same individual can vary greatly, causing inconsistency in the measurement. West et al. (2018) suggested that the complex nature of the procedural processes, relative to processes related to declarative memory, is the reason why tasks used for assessing procedural learning (i.e., the serial RT tasks) were significantly less reliable than the tasks for declarative learning (i.e., free recall tasks) in their data. In other cases, researchers believe that some of these tasks are not measuring what it is designed to after all; instead, they might be tapping into theoretically less interesting constructs, such as processing speed and strategies (e.g., Hedge et al., 2022; Miller & Ulrich, 2013; Rouder et al., 2022).

The second source of unreliability can be statistical. Specifically, the low reliability can be due to a lack of sufficient variance between participants (e.g., Clark et al., 2022; Hedge et al., 2018). This means that participants are not different enough to show reliable individual differences. This may be due to homogeneous sampling and the nature of the attribute being measured in which individuals do not differ much (Hedge et al., 2018). This can also be caused by the design of the task. For example, it can be too easy for the sample and thus, there is a ceiling effect; and/or it can be that there is little variability in item difficulty (Clark et al., 2022; Hedge et al., 2018). Related to task design, intuitively, data might have some random noise if the participant is tested from home and the set-up of the participant's technology plays a critical role (e.g., Patterson & Nicklin, 2023).

Finally, the source of unreliability can also be computational. That is, the way in which the RT-difference data are preprocessed and analyzed could have contributed to the low level of consistency. For example, Buffington et al. (2021) speculated that computing difference scores between trial types may be "the source of low reliability" (p. 647). In the analysis code that these authors shared, they applaudingly documented their thought process in pinpointing the cause of inconsistency. They wrote that outliers "needed to be addressed," and the outlier treatment strategy they had employed "did seem to help." Among all potential sources, the computation of consistency levels

https://doi.org/10.1017/S027226312300027X Published online by Cambridge University Press

is possibly the one that applied researchers, who use these measures to address their substantive research questions (as opposed to methodologists who examine these methods), can act on because they can compute reliability based on an informed approach. If that is the case, the natural question is then how researchers should estimate reliability for RT-difference data.

Estimating reliability for response-time differences

As mentioned, there is a lack of consensus on how to estimate internal consistency for RT-difference data. Here, we discuss three methods: computing RT differences based on (1) raw RTs, (2) by-participant *z*-transformed RTs, and (3) estimates of RT differences based on mixed-effects modeling.

First, an intuitive approach is to compute the RT difference for each item across two trial types (e.g., related vs. unrelated or grammatical vs. ungrammatical). For example, when a slowdown in processing is expected for grammatical violations, a 500-ms response on a grammatical trial and a 550-ms response on an ungrammatical trial translate into a 50-ms slowdown. This difference can be used to index one's grammatical sensitivity. What Buffington et al. (2021) argued is that this difference score, aggregated across items, is not reliable. The unreliable nature of difference (or change) scores has long been a concern in social sciences (e.g., Cronbach & Furby, 1970; Gulliksen, 1950). Similar arguments have been made more recently in the context of RT research (Draheim et al., 2019). One reason for such an inconsistency is that subtraction can reduce the between-participant variance relative to error variance (e.g., Hedge et al., 2018). In other words, by subtracting the RT in one trial from that in another trial, the researcher removes the useful, common information carried by the two RTs that makes the individual participant unique in the sample (i.e., the between-participants variance). What is left then is random variation within the individual (i.e., within-participant variance), contributing to the overall unreliability. Another criticism of a raw difference is that it does not adequately account for baseline differences between individuals (e.g., Tan & Yap, 2016). For example, a 50-ms slowdown indexes somewhat different levels of change for a learner whose baseline RT was 500 ms versus their peer whose baseline was, say, 300 ms. In this respect, taking one's baseline into account could yield better reliability.

One way to do that is the second approach that we discuss here. Following this approach, researchers first compute *z*-score-transformed RTs, based on each participant's own mean and standard deviation (Hutchison et al., 2008; Tan & Yap, 2016). This step allows researchers to put participants on an equal footing because all participants have a mean RT of zero and the change scores (RT differences) are expressed in their own standard deviation unit. In this way, the baseline RT for each participant is accounted for in the computation, which addresses the limitations of using raw RTs.

However, there are two additional points to note here: First, following either of these approaches, researchers may need to decide on their handling of missing data. The RTs can be coded as missing when participants respond incorrectly or when the RT is outside the data-trimming threshold. Indeed, (applied) psycholinguists often adopt a counterbalanced design where each participant sees only one of two versions (e.g., grammatical) of the item, in order to avoid participants being exposed to very similar trials and thus creating unwanted confounding. All these situations are not uncommon in the analysis of data in psycholinguistics. One way to handle missing data in estimating reliability is thento discard the item when one of the two RTs is missing. This can result in discarding more data than necessary and has implications for

https://doi.org/10.1017/S027226312300027X Published online by Cambridge University Press

statistical power and the intended inference researchers wish to make. The second way to get around the problem is to perform an aggregation before subtraction. That is, each participant has a mean RT for each of the two trial types before the computation of an RT difference. If there were no missing data, the two orders of operation (i.e., aggregation before subtraction and subtraction before aggregation) yield identical results mathematically. However, with missing data, aggregation before subtraction means that average RTs likely result from different items. Although averaging across (a large number of) items should generally lead to more accurate results, one might still question the extent to which the items that go into the computation for both trial types are similar enough to warrant a direct comparison.

Second, aggregating an effect across items, be it before or after subtraction, should remind applied psycholinguists of how researchers used to carry out separate by-participant and by-item analyses for RT data, which are no longer recommended (e.g., Baayen et al., 2008). The reason is that by-participant aggregations essentially ignore the variability in the effects associated with the item and vice versa. The more contemporary approach is to simultaneously model participant and item variability by implementing a mixed-effects model (e.g., Baayen et al., 2008). Generally, a mixed-effects model results in more accurate estimates and represents a more parsimonious analysis of the data. It is not an exaggeration to suggest that applied psycholinguists are already intimately familiar with the technique.

In the present context of estimating the reliability for RT differences, the use of such a model is rare in SLA. This approach represents the third approach we discuss here: the model-based approach. Not adopting this approach can be seen as a missed opportunity because, again, many researchers are already familiar with these types of models. Moreover, it has the advantage of simultaneously modeling random effects associated with both item and participant, accounting for dependency in the data as a result of individual-difference factors specific to the participant (e.g., processing speed) and characteristics specific to the item (e.g., frequency). Perhaps less discussed in SLA is also the ability of a mixed-effects model to handle missing data through (restricted) maximum likelihood (e.g., Hox et al., 2018), addressing the missing data challenge discussed above. Therefore, a model-based approach should be a promising candidate to arrive at more accurate reliability estimates.

One way to understand a mixed-effects model is to imagine that a regression line is fitted for each participant and for each item (level-2 units). That means that every level-2 unit has its own intercept and slope terms (when they are allowed to vary). When there are, for example, 40 participants in the data set, there can be 40 intercept and 40 slope values, as well as a correlation between them. Simultaneously, the fixed effects, which researchers often interpret, are computed by the algorithm, taking into account these random effects. The critical information here in relation to reliability assessment is this: The by-participant random slope for each individual represents a model-based summary of the main effect (slowdown or speedup) specific to the learner. Therefore, the by-participant random slopes can be seen as individualized difference scores, *after* accounting for all relevant random effects. The use of by-participant random-slope values as a basis for estimating reliability, on this account, should produce better results.

We should note, at this point, that we are not the first to suggest that mixed-effects models can be used to estimate the reliability of RT-difference measures. Previous authors (e.g., Rouder & Haaf, 2019) have already suggested the same. Rouder and Haaf (2019) were interested in the cause of the low correlation between two attentional control tasks, the Stroop and flanker tasks, as reported by Hedge et al. (2018). They retested the data of Hedge et al. (2018) and reported better test-retest correlations based

https://doi.org/10.1017/S027226312300027X Published online by Cambridge University Press

on model estimates. Overall, the authors observed an increase of around .20, compared with the non-model-based sample correlations. Given the promising evidence of Rouder and Haaf (2019), it is high time to test the extent to which a model-based approach to estimating reliability is appropriate for the RT-based L2 data.

In addition, the model-based approach offers an unexplored opportunity to treat outliers in reliability assessments. As discussed, Buffington et al. (2021) pointed out in their analysis code that outliers can be detrimental to the overall instrument reliability. Although the authors handled outlier RTs using more conventional strategies based on means and standard deviations of individual participants, the extent to which a model-based approach might offer better results remains an open question. In RT-based research, Baayen and Milin (2010) have shown the benefits of engaging in model criticism as a way to trim RT data, although model criticism is not always used to handle outliers in analyzing SPR data (e.g., Marsden et al., 2018). This approach amounts to fitting an initial mixed-effects model to the raw data to first identify and remove observations with a large, standardized residual (e.g., an absolute value larger than 2.5, Baayen & Milin, 2010). With the trimmed data set, researchers refit the model to the data with the same model specifications. According to these authors, the refitted model almost always has a better fit, with fewer observation removals needed. On this basis, the reliability resulting from this procedure should represent a more accurate estimate because the better model fit should produce more accurate by-participant random slopes. Therefore, the application of this technique could further improve the estimated reliability of the RT differences under investigation.

The present studies

Considering this review, we formulated two research questions for our first study to assess the model-based approach to estimating reliability:

> RQ1: To what extent does a model-based approach yield more reliable RT differences?
> RQ2: To what extent does model criticism as an outlier treatment strategy yield more reliable RT differences?

Based on the results of the first study, we further addressed RQ3 in our second simulation study to examine the boundaries of the model-based approach:

> RQ3: Under what conditions, in terms of the number of items and level of error, is a model-based approach more beneficial in estimating the reliability of RT differences, than non-model-based approaches?

Taken together, our current attempt represents an important and ethical step for evaluating RT-difference measures that SLA researchers begin to increasingly rely on, moving beyond simply accepting the face value of our instruments without scrutinizing (much) their reliability and validity (Cohen & Macaro, 2013).

Study 1: Analysis of three open data sets

Methodology

To address our first two research questions, we took advantage of open data shared by L2 researchers. Three data sets, including data from an SPR task, a lexical decision task

https://doi.org/10.1017/S027226312300027X Published online by Cambridge University Press

in masked-repetition priming, and an alternating serial reaction time task, were used to estimate the reliability of RT differences. For RQ1, we tested three computational approaches: RT differences based on (1) raw RTs, (2) by-participant *z*-transformed RTs, and (3) model-based estimates of RTs. Addressing RQ2, we compared trimming strategies based on mean and standard deviations (following the initial authors of the data sets) with the implementation of model criticism. In the spirit of open science, all R code (R Core Team, 2022) used for data analysis is made available in the Open Science Framework (https://osf.io/cd5r8/).

Data sets

Here, we provide minimal background information to understand the contexts from which the original data were collected. Interested readers are referred to the substantive publications associated with the data sets.

The first data set was shared by Fang and Wu (2022a), available on Open Science Framework (https://osf.io/abhjv/). The associated substantive publication was Fang and Wu (2022b). The authors administered SPR and acceptability judgment tasks to investigate learners' (N = 122; 135 initially, 13 removed) knowledge of the *either–or* construction in English. We used only the SPR data involving learners (bind_SPR_English_L2_version1-2.csv). Participants read, in a self-paced manner, two versions of 20 sentences: one with *either* (e.g., *Jay painted either | the big house | or the old car | for his family over the summer*) and one without (e.g., *Jay painted | the big house | or the old car | for his family over the summer*). The authors reported a significant speedup in reading times in the critical region (i.e., *or the old car*) on the trials that included *either* compared with those without, suggesting that learners can use knowledge of the *either–or* construction to make predictions in reading. Given that the claims made by the researchers were largely based on a speed up in the critical region, we also focused on this region in our analysis as well.

The second data set we used was made public by Hui et al. (2022b), available on Open Science Framework (https://osf.io/uyfh5/) under the Creative Commons Attribution 4.0 International Public License. The associated substantive publication was Hui et al. (2022a). In their study, the authors administered a set of four vocabulary tests to 129 (144 initially, 15 removed) advanced English learners at an American university. We used only the subset of data for masked-repetition priming (i.e., data_exp_35094-v18_task-uoc2.csv). As discussed, this task was used as a vocabulary measure to index the extent to which the lexical entries of a sample of target words (K = 40) had been established in the mental lexicon. In the 80 critical trials (two for each of 40 items), participants were exposed to a prime presented very briefly (55 ms) and forward masked by a string of hashtags (####) for 500 ms. Immediately after, the participant made a lexical decision on the target presented in the upper case to indicate whether or not it forms an English word. The authors reported a group-level priming effect in which participants responded faster to the target in the related, identical trials (e.g., patience-PATIENCE) than to the unrelated trials (e.g., occasion-PATIENCE). The authors also reported that such priming was not observed for their nonword data, providing further evidence for the validity of the measure.

The third data set was shared by Buffington and Morgan-Short (2022), available on Open Science Framework (https://osf.io/ux4qs/). The associated substantive publication was Buffington et al. (2021). The authors performed a total of six memory assessments (three for procedural memory and three for declarative memory). We used only the data set for the alternating serial reaction time task (i.e., ASRT_MasterData.csv). This task

https://doi.org/10.1017/S027226312300027X Published online by Cambridge University Press

tested the ability of the participants (N = 99; 119 initially, 20 removed) to acquire an implicit, patterned sequence in the task, which is associated with the use of procedural memory. As reviewed, participants pressed a button corresponding to the location of a dog's head appearing in one of four circles. The sequence followed a second-order pattern in that the patterned trials alternated with the random trials. There were a total of 20 experimental blocks, each of which had 85 trials (five random trials to start, followed by 80 alternating patterned and random trials).

Analysis

For each data set, we computed a total of 22 split-half correlations, following the three computational approaches with and without a trimming procedure (RQ1). In the case of the model-based approach, we tested an additional trimming method—namely, model criticism (RQ2). For the first two computational approaches (raw and z scores), we implemented both by-participant and by-item analyses. For all reliability estimates, we performed two correlation tests (Pearson and robust) for each analysis method (see Figure 1).

To analyze the data, we first repeated all accuracy-based screening procedures following the original authors. We also removed practically impossible RTs, as defined by the authors (negative RTs for Fang & Wu, 2022a; 300 ms for Hui et al., 2022b; 100 ms for Buffington & Morgan-Short, 2022). This treatment was to remove completely unusable data and differed from the trimming procedure that seeks to rid the data sets of outliers. Although Fang and Wu (2022b) logarithmically transformed and residualized the RTs on the length of the region before their further analysis, we used the "raw" RTs in our analysis for consistency across the three data sets. As a sensitivity analysis, we confirmed that using the logged, residualized RTs did not change our conclusion. The resulting data sets from these preliminary processing steps represented the untrimmed data sets for further computation.

To create the trimmed data sets, we also deleted RTs that were above the authors' upper threshold (2.5 standard deviations from the learner's mean for Fang & Wu, 2022a; 2500 ms for Hui et al., 2022b; and 3.0 standard deviations from the participant's mean for Buffington et al., 2021).

As a next step, we split each data set into two halves (i.e., odd-numbered and even-numbered items). For Buffington and Morgan-Short (2022), we needed to assign item

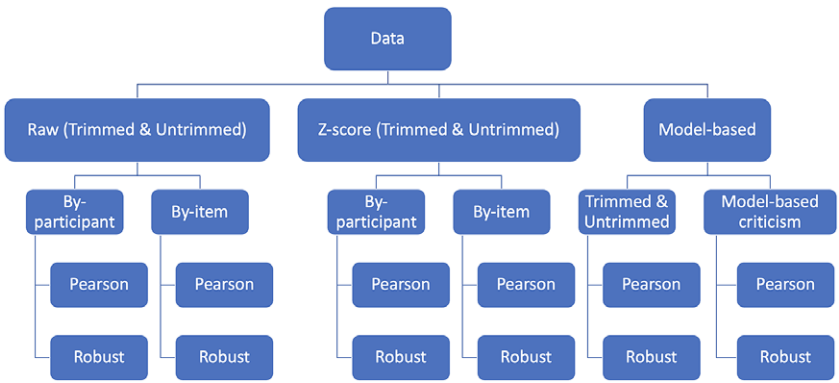

Figure 1. The three computational approaches and the corresponding calculations.

https://doi.org/10.1017/S027226312300027X Published online by Cambridge University Press

numbers to alternating trials so that every two back-to-back trials (one patterned and one random) was considered a duplet item. Thus, we addressed issues associated with learning that takes place during the task that can lead to serial dependence among trials (see more discussion in Buffington et al., 2021).

To take our first computation approach, we calculated the raw difference in RTs for each item. We aggregated it across both participants and items, such that each item and each participant have a mean RT difference, respectively. To estimate split-half reliability, we computed both Pearson and robust correlations with the two halved data sets, using the cor.test( ) function in the stats package (R Core Team, 2022) and the pbcor( ) function in the WRS2 package (Mair & Wilcox, 2020).

For the second approach, we transformed the RTs into z scores based on each participant's own mean and standard deviation. After that, we computed the difference and aggregated it across items and between participants. The correlation tests were performed the same as in the first approach.

In terms of the model-based approach, we constructed two separate models for each of the two halved data sets. For all three data sets, the outcome was specified as the inverse of RT (-1/RT), following Hui et al. (2022a). We used a maximal random-effects structure, including all relevant random intercepts and slopes, as well as their correlation (Barr et al., 2013). To resolve issues with singular fit and nonconvergence, we used the *nloptwrap* optimizer from the optimx package (Nash & Varadhan, 2011) and the partial Bayesian method implemented in the blme package (Chung et al., 2013) to force the relevant random-effects matrices away from singularity. From the models, we extracted the by-participant random slope for each participant before we correlated the slopes across the two halves of each data set.

Finally, we tested whether and how model criticism as an outlier treatment strategy might be useful (RQ2). To engage in model criticism, we first used the residuals of the models fitted to the untrimmed data as an outlier identification strategy and then removed observations that had a standardized absolute residual greater than 2.5 (Baayen & Milin, 2010). We refitted the models to these trimmed data sets using the same specifications and extracted the by-participant random slopes for the correlation tests.

As an additional note, whenever a correlation test returned an unexpected negative value, we followed Buffington et al. (2021) in applying a correction according to Krus and Helmstadter's (1993) Equation 15—namely, $corrected\ r = \frac{-r_{ab}}{.5(1-r_{ab})}$. We have marked the cases where the correction was applied in our results.

Results

We present the correlation coefficients in Tables 1 to 3. Before we address our research questions, one striking observation is that the correlation coefficient can vary hugely depending on the data analysis undertaken. With the same data set (e.g., Hui et al., 2022b), the coefficient ranged from .02, indicating essentially no association between what was measured in the two halved data sets, to .88, a satisfactory level of reliability. This variability confirmed the dire need to develop a more informed and standardized approach to estimating reliability.

In terms of RQ1, concerning the usefulness of the model-based approach, only the results for the Hui et al. (2022b) data showed a superiority of the model-based approach over the other two approaches. The reliability was in the .30 range at best with non-model-based approaches. However, when the by-participant random slope was used as a basis for reliability estimation, the figures were mostly in the .80 range, suggesting

https://doi.org/10.1017/S027226312300027X Published online by Cambridge University Press

Table 1. Split-half correlations for the Fang and Wu data set

	Pearson correlation coefficient			Robust correlation coefficient		
	Trimmed	Untrimmed	Model criticism	Trimmed	Untrimmed	Model criticism
Raw response times–by participant	.36	.16	NA	.36	.05	NA
Raw response times–by item	.10	.29#	NA	.02	.32#	NA
By-participant *z*-transformed response times–by participant	.15#	.10#	NA	.17#	.06#	NA
By-participant *z*-transformed response times–by item	.15	.36	NA	.15	.36	NA
Model-based estimates	.16#	.20#	.02#	.42#	.39#	.16#

Note: # indicates that a correction was applied to a negative coefficient according to Krus and Helmstadter (1993). We do not report 95% CIs and *p* values because it is not clear whether a correction for these statistics is needed and, if so, how to compute them.

Table 2. Split-half correlations for the Hui et al. data set

	Pearson correlation coefficient			Robust correlation coefficient		
	Trimmed	Untrimmed	Model criticism	Trimmed	Untrimmed	Model criticism
Raw response times–by participant	.11#	.07	NA	.17#	.08	NA
Raw response times–by item	.36	.11#	NA	.23	.04#	NA
By-participant *z*-transformed response times–by participant	.03	.15	NA	.01	.15	NA
By-participant *z*-transformed response times–by item	.02	.15	NA	.11#	.04	NA
Model-based estimates	.85	.81	.88	.81	.79	.87

Note: # indicates that a correction was applied to a negative coefficient according to Krus and Helmstadter (1993). We do not report 95% CIs and *p* values because it is not clear whether a correction for these statistics is needed and, if so, how to compute them.

Table 3. Split-half correlations for the Buffington and Morgan-Short data set

	Pearson correlation coefficient			Robust correlation coefficient		
	Trimmed	Untrimmed	Model criticism	Trimmed	Untrimmed	Model criticism
Raw response times–by participant	.47#	.06	NA	.51#	.44#	NA
Raw response times–by item	.02	.01	NA	.01#	.05	NA
By-participant *z*-transformed response times–by participant	.51#	.48#	NA	.47#	.46#	NA
By-participant *z*-transformed response times–by item	.07	.09	NA	.06	.10	NA
Model-based estimates	.46#	.31	.50#	.30	.30	.49#

Note: # indicates that a correction was applied to a negative coefficient according to Krus and Helmstadter (1993). We do not report 95% CIs and *p* values because it is not clear whether a correction for these statistics is needed and, if so, how to compute them.

https://doi.org/10.1017/S027226312300027X Published online by Cambridge University Press

satisfactory levels of internal consistency. In contrast, neither the Fang and Wu (2022a) data nor the Buffington and Morgan-Short (2022) data presented a clear pattern of which analysis method had the greater advantage in reliability. The relatively higher levels following the model-based approach could also be achieved using other approaches.

For RQ2, we focused on comparing the correlations computed from the trimmed data (using conventional strategies) and from model criticism. Model criticism as an outlier treatment strategy appeared to have improved the figures of both the Hui et al. (2022b) and the Buffington and Morgan-Short (2022) data sets. In the case of the Hui et al. data, the improvement was not as marked, potentially due to the already high level of correlations (.81 and .85). No advantage was observed for the Fang and Wu data.

Discussion

In our analysis of the three open data sets, we found that the model-based approach was useful to varying degrees. Overall, the model-based approach only showed a clear advantage for the Hui et al. (2022b) data set. This was, at least initially, surprising to us. Here, we first discuss why the model-based approach demonstrated exceptional performance with this data set. Then, we consider why this advantage did not generalize to the other two data sets, which motivated our simulation study that follows.

Repeatedly, we have stressed that mixed-effects models can simultaneously model variability caused by characteristics of *both* individual items and participants (Baayen et al., 2008). Therefore, the benefits of the mixed-effects approach should be most obvious when there is a reasonable amount of participant and item variability to be accounted for. In the Hui et al. (2022b) data, the authors used the masked-repetition priming as a vocabulary test. Therefore, they sampled word stimuli across four frequency bands. Put differently, words included in the experiment were of various levels of difficulty. Highly frequent words (e.g., *upset*) should elicit much faster responses than items at a lower frequency band (e.g., *miniature*). This means that the intercept terms, representing the baseline RT for each item, can vary to a large extent. Relatedly, the level of priming might also vary because there is not a lot of room for a drastic speedup if the baseline RT is already fast. Therefore, properly modeling item variability (in addition to participant variability) has proven to be the right strategy because the accuracy of the estimates was improved through a phenomenon known as shrinkage or regularization, a primary property of mixed-effects models (Baayen et al., 2008; Winter, 2019). More accurate estimates also mean that the person-related parameters are closer to their true values (e.g., baseline RTs [intercept] and priming effects [slope]); therefore, the improvement in the split-half reliability was especially obvious for this data set.

In contrast, the Buffington and Morgan-Short (2022) and Fang and Wu (2022a) data sets did not have a large amount of item variability. In Buffington and Morgan-Short (2022), there was little variability in terms of difficulty because there were consistently four positions in which the dog head could appear. The same was true for the study by Fang and Wu (2022b), who focused only on one single grammatical structure (i.e., the *either–or* construction). Perhaps, when there is little item variability to be accounted for in the first place, the design of the task might prevent the mixed-effects models from achieving their full potential. In addition, the number of items might have played a role. In the Buffington and Morgan-Short data, there were 800 items (or 1,600 trials). Aggregating across such a large number of items yields very accurate results, as

https://doi.org/10.1017/S027226312300027X Published online by Cambridge University Press

manifested by the higher levels of reliability in the by-participant analysis than in the by-item analysis (see Table 3). Therefore, the aggregation-and-subtraction approaches were not disadvantaged due to the sample size of the items. Finally, the overall level of error should be a determining factor because the model-based approach is not a magic wand that can remove all error and make an unreliable measure suddenly reliable. What mixed-effects models are capable of is properly partitioning variance, thus accounting for variance due to participant and item that could have been regarded as error variance contributing to the unreliability of the measure. However, it cannot remove random error from the data set.

In light of these accounts, we performed a simulation study to show the effects of (1) item number and (2) level of error on the usefulness of the model-based approach to estimating reliability for RT differences (RQ3).

Study 2: Simulations

In the previous analyses, we observed that only the Hui et al. (2022b) data benefited from a model-based approach to estimating reliability. The overall objective of this simulation study was to explore the strengths and limitations of the model-based approach. It is also through this study that we address our RQ3, which we repeat here for easier reference:

> RQ3: Under what conditions, in terms of the number of items and level of error, is a model-based approach more beneficial in estimating the reliability of RT differences, than non-model-based approaches?

Methodology

Data simulation

Data simulation refers to creating artificial data sets for analysis. By definition, simulated data are not authentic in the sense that they are not collected from human participants. However, because researchers have the flexibility to determine the characteristics of the data in the data generation process and a large number of data sets can be created in one go to simulate various scenarios, it has been a useful tool for methods research in fields such as psychology and educational sciences. Because methods research in SLA is only taking off, the use of simulations is not yet common. Among a small number of exceptions, power analysis appears to be one of the main uses of simulations (e.g., Brysbaert & Stevens, 2018; Nicklin & Vitta, 2021; Vitta et al., 2021). By manipulating parameters such as effect sizes in artificial data sets, followed by implementing the relevant statistical test, the authors were able to provide sample-size guidelines for applied researchers to achieve sufficient levels of statistical power. In addition, data simulation has been used in methods education, allowing researchers to more fully understand how certain statistical models operate without the need to deal with the contamination of uncertainty naturally embedded in real data sets from research studies (DeBruine & Barr, 2021). In the present analysis, we simulated data sets, followed by reliability assessments, in order to identify any "sweet spots" where the model-based approach would be most helpful in estimating reliability.

The simulation

We wrote our R code (available on our OSF page: https://osf.io/cd5r8/), based on DeBruine & Barr (2021) and an online tutorial provided by DeBruine (2020).

https://doi.org/10.1017/S027226312300027X Published online by Cambridge University Press

Table 4. Parameters for the baseline data set

R code variable	Data specification	Value
sub_n	number of participants	120
sub_sd	standard deviation (SD) for participants' random intercepts	0.29
sub_version_sd	participants' slopes SD	0.03
sub_i_version_cor	participant intercept-slope correlation	−0.97
stim_n	number of stimuli in the simulation	40
stim_sd	SD for the stimuli's random intercept	0.08
stim_version_sd	stimuli's slope SD	0.03
stim_i_cor	correlations between intercept and slopes	−0.18
grand_i	overall mean of the dependent variable (DV)	−1.50
stim_version_eff	mean difference between versions: related - unrelated	0.10
error_sd	residual (error) SD	0.20

As a first step, we created a baseline data set whose parameters were set up based on the model summary reported by Hui et al. (2022b; see Table 4), because the model-based approach was proven the most useful for the Hui et al. (2022b) data set as reported in Study 1. In other words, the parameters were specified according to the model reported in Hui et al. (2022b). In brief, we simulated a situation with a within-participant design where participants ($N = 120$) responded to 40 items containing both related and unrelated trials. The mean difference between versions was set at 0.10, given a grand mean of -1.50. With this simulated data set, we repeated two reliability assessments, following the steps reported in Study 1 above. One assessment was based on RT differences from raw RTs, and the other was based on the model-based approach. We did not perform the *z*-transformation procedure as in Study 1 because the results were not promising in that it did not produce higher levels of reliability in the three open data sets analyzed.

Addressing RQ3, we tested the effects of varying two parameters, the number of items and the degree of error, while keeping other specifications constant. We then compared the performance of the two approaches on reliability. For the number of items, we tested eight levels (i.e., $k = 20, 30, 40$ [baseline], 80, 120, 160, 400, and 600). In terms of the degree of error, we tested seven levels (i.e., Residual $SD = 0.05, 0.10, 0.20$ [baseline], 0.40, 0.60, 0.80, and 1.00).

Results

We present and visualize the correlation coefficients as well as their 95% confidence intervals in Tables 5 and 6 and in Figures 2 and 3. Overall, increasing the number of items improves the reliability for RT differences based on raw RTs. The model-based approach reached its ceiling levels when there were 30 to 40 items. In contrast, the raw RT approach needed up to 400 items to reach acceptable levels. For the degree of error, increasing the level of noise in the data caused the reliability to drop, which was expected. However, it appeared that the drop was more drastic for the raw RT approach than for the model-based approach. When we doubled the levels of error (i.e., from 0.20 to 0.40), the split-half correlation for the raw RT approach essentially floored at .02, compared with a satisfactory level of .82. At the same time, a further increase in error resulted in a deeper dive, even for the model-based approach.

https://doi.org/10.1017/S027226312300027X Published online by Cambridge University Press

Table 5. Split-half correlations for simulated data sets (varying the numbers of stimuli)

Number of stimuli	Raw RT reliability [95% CI]	Model-based
20	.03 [−.15, .21]	.79 [.71, .85]
30	.12 [−.06, .29]	.90 [.86, .93]
40 (baseline)	.23 [.05, .39]	.95 [.93, .97]
80	.32 [.15, .47]	.95 [.93, .97]
120	.35 [.18, .50]	.98 [.97, .99]
160	.50 [.35, .62]	.98 [.98, .99]
400	.69 [.59, .78]	.98 [.97, .99]
600	.79 [.72, .85]	.99 [.99, .99]

Table 6. Split-half correlations for simulated data sets (varying the degrees of error)

Error *SD*	Raw RT reliability [95% CI]	Model-based
0.05	.81 [.73, .86]	.99 [.98, .99]
0.10	.52 [.38, .64]	.99 [.98, .99]
0.20 (baseline)	.23 [.05, .39]	.95 [.93, .97]
0.40	.07 [−.11, .25]	.82[.76, .87]
0.60	.03 [−.15, .21]	.63 [.50, .72]
0.80	.02 [−.16, .19]	.39 [.23, .53]
1.00	.01 [−.17, .19]	.18 [.00, .35]

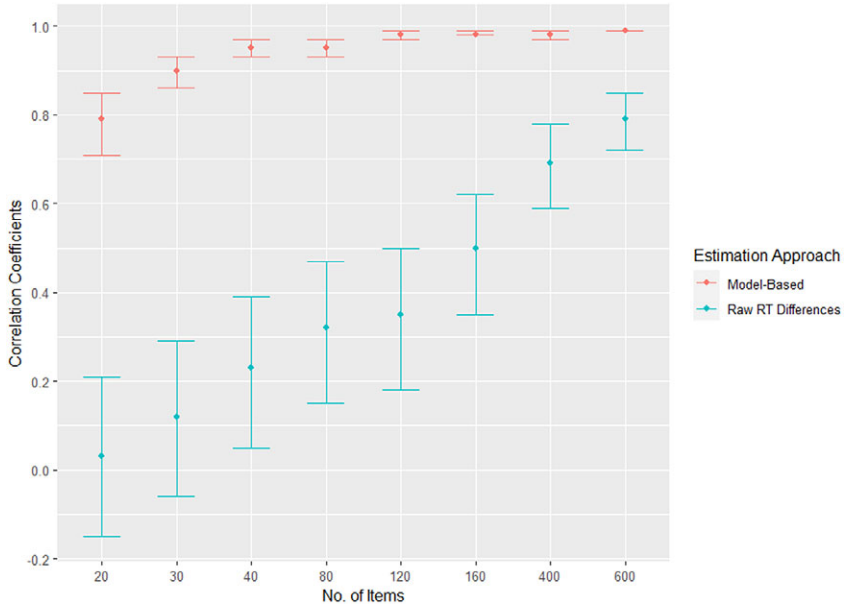

Figure 2. Correlations and their confidence intervals for two estimation approaches varying the number of items.

Discussion

Through a series of simulations, we have shown that the model-based approach can be a promising computational alternative in estimating the error of RT-difference measures.

https://doi.org/10.1017/S027226312300027X Published online by Cambridge University Press

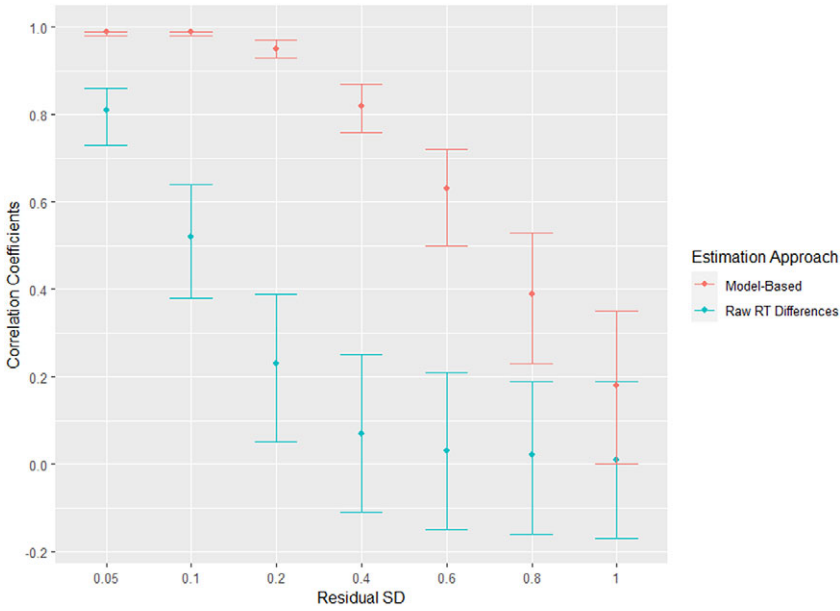

Figure 3. Correlations and their confidence intervals for two estimation approaches varying the level of error.

It appears to be particularly useful when the researchers do not have a large number of items (e.g., $k = 400$) and when the level of error is only moderate. In contrast, the raw RT approach can also be useful when the number of items reaches a certain threshold and when the error standard deviation—that is, the level of noise—is small enough. Under these specific circumstances, the averaging of raw RT differences can also be powerful enough that a model-based approach becomes unnecessary. In other more extreme conditions, such as when the error level is too high, neither approach can provide a good rescue. Taken together, the simulation confirms our intuition that there can be a "sweet spot" where the model-based approach can be most useful. From the present analyses, this could mean it is advisable to apply a model-based approach in situations where the researchers cannot include a large number of items and the level of error is only moderate.

General discussion and conclusion

In this article, we took a step to search for a more informed, potentially standard computational method to estimate the reliability of RT differences, a type of measure that SLA researchers are beginning to rely on. We carried out two sets of analyses using both open, authentic (Study 1) and artificial, simulated (Study 2) data sets. The results of Study 1 demonstrated that model-based approaches can be very powerful in yielding a high reliability level compared with methods based on raw RT or z-transformed RT differences, but only in certain cases, and that adding the model criticism did not yield more reliable results. In Study 2, we highlighted situations where the model-based approach can be useful, but we also demonstrated that this approach is not a magic wand that can remove error from an unreliable measure. Therefore, we argue that

https://doi.org/10.1017/S027226312300027X Published online by Cambridge University Press

the model-based approach presented in this article represents an alternative that researchers should consider, especially under certain circumstances. As in any data analysis, there is almost always more than one acceptable procedure (see, e.g., Steegen et al., 2016). For example, Suzuki et al. (2022) reported Cronbach's alpha for their word-monitoring measure (as a grammatical sensitivity index). Researchers have also examined the use of a model-based Bayesian approach on RT-based tasks (Haines et al., 2020). The present analysis remains silent on how our approach, based on mixed-effects modeling, would compare to theirs.

When the model-based approach yields a good level of reliability, researchers should use it. In that case, provided that the RT difference is the outcome, researchers should engage in mixed-effects modeling, as they probably would. On the other hand, if the RT difference is on the predictor side of the equation, as an individual-difference measure, we suggest that researchers first model the RT data with a mixed-effects model. Then, they can use the by-participant random slopes for the condition as a predictor and/or allow it to interact with other variables. In work involving structural equation modeling, such as the measurement of explicit and implicit knowledge, the RT-difference measure may be an indicator of a common factor. The use of the random slopes in these cases is also appropriate because of the relatively high reliability (although measurement error is already taken into account in the building of a common factor model).

However, we must stress that researchers should treat study design as the first line of defense in tackling potential unreliability. For example, to the extent that resources allow, researchers may consider having a greater number of items. Brysbaert and Stevens (2018), for example, recommended having 1,600 observations (e.g., 40 participants times, 40 items) to run a well-powered study. When considering sample-size planning, many SLA researchers focus almost entirely on the interplay between the number of participants, expected effect sizes, significance levels, and statistical power (see, e.g., Loewen & Hui, 2021). In fact, the number of items (shown in Study 2) and the reliability of the instrument (discussed in the introductory sections) should receive more attention. In addition to the number of items, there should ideally be sufficient variability in terms of item difficulty so as to differentiate able and less able participants. This can be achieved by including more grammatical structures or words in different frequency bands. For example, Suzuki et al. (2022) and Godfroid and Kim (2021) had four and six grammatical structures in their test, respectively. Hui et al. (2022b) included vocabulary items in four frequency bands. All these design features will help researchers reach a satisfactory level of reliability.

More generally, we echo previous calls for more consistent reporting of instrument reliability (Marsden et al., 2018; Plonsky & Derrick, 2016). In the present case of RT-difference measures, researchers can report the split-half reliability based on the by-participant random slopes for experimental conditions. This information is critically important for understanding the uncertainty surrounding the measurement and the limitations of the instrument. Consumers of research need this information to evaluate claims made in studies and/or to determine the extent to which the instrument is of sufficient quality to be adopted in subsequent research. As L2 researchers increasingly embrace open science practices, more materials are publicly shared in repositories such as IRIS (Marsden et al., 2016). Reliability information represents one of the criteria that researchers can take advantage of to make an informed assessment of the quality of the instrument.

In relation to adopting materials, researchers should also share trial-level data (e.g., Hui et al., 2023; Isbell, 2021; Marsden & Morgan-Short, 2023). Although the overall reliability is informative, subsequent researchers are in a position to further improve the

https://doi.org/10.1017/S027226312300027X Published online by Cambridge University Press

instrument in their own study through an item analysis of the shared data. For example, specific items may not be eliciting the intended effects (e.g., priming) due to some overlooked item characteristics. These incidences can create noise in the data that can be difficult to identify by simply inspecting the overall reliability levels. Access to trial-level data allows subsequent researchers to revisit and inspect the items more thoroughly (e.g., by examining the by-item random slopes) to identify potentially questionable items. Then, they can either revise or remove the item(s), making the overall instrument a better version in subsequent work. We would like to mention that our search for appropriate data sets to analyze for Study 1 was not straightforward. For example, we encountered studies that were awarded an Open Data badge, but trial-level data were not included. Despite the challenges, we are very thankful for researchers who do share their data because they have made more methods research possible.

Moreover, we join previous calls for the sharing of reproducible analysis code (Hui & Huntley, 2021; In'nami et al., 2022). Analysis code can reveal the procedure one has undertaken to arrive at the reported reliability. As discussed in the introductory sections, specific steps in the estimation of reliability for RT differences may be seen as a mystery because the computation is often not detailed and there is no standard practice. In such cases, many applied researchers would benefit from knowing how reliability is computed in previous studies. That is not to say that researchers should blindly follow what others have done. On the contrary, the analysis code provides methodologists with a way to understand current practices, a first step toward refining them.

Finally, although the scope of this study is limited to reliability, we would like to raise researchers' awareness of validity issues because the ultimate goal for research is to have both reliable and valid measures. Reliability is only a necessary, not sufficient, criterion for a valid measure. Even when there are ways to promote reliability, these tasks can still be reliably measuring an irrelevant construct. If that turns out to be the case, SLA researchers might want to consider redesigning and/or developing new measures intended specifically for individual differences research (Burgoyne et al., 2022; Draheim et al., 2022; Draheim et al., 2021; Weigard et al., 2021).

To conclude, we have demonstrated that mixed-effects modeling can provide an alternative for researchers to estimate the reliability of RT differences more accurately and that model criticism as an outlier treatment strategy does not necessarily result in better reliability and thus may not be necessary for measuring reliability. Although this approach performs very well under specific conditions and it cannot magically remove error from a measure, it represents a promising option for researchers using RT differences in their work. We also hope that this article can enhance researchers' awareness of the need to constantly evaluate the psychometric properties of the measures we rely on.

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

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. https://www.testingstandards.net/uploads/7/6/6/4/76643089/standards_2014edition.pdf

Arnon, I. (2019). Do current statistical learning tasks capture stable individual differences in children? An investigation of task reliability across modalities. *Behavioral Research Methods*, *52*, 68–81. https://doi.org/10.3758/s13428-019-01205-5

https://doi.org/10.1017/S027226312300027X Published online by Cambridge University Press

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*, 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Baayen, R. H., & Milin, P. (2010). Analyzing reaction times. *International Journal of Psychological Research*, *3*, 12–28. https://doi.org/10.21500/20112084.807

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*, 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bowles, M. A. (2011). Measuring implicit and explicit linguistic knowledge: What can heritage language learners contribute? *Studies in Second Language Acquisition*, *33*, 247–271. https://doi.org/10.1017/S0272263110000756

Brill-Schuetz, K., & Morgan-Short, K. (2014). The role of procedural memory in adult second language acquisition. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *36*, 260–265. https://escholarship.org/content/qt0dc7958r/qt0dc7958r.pdf

Brysbaert, M., & Stevens, M. (2018). Power analysis and effect size in mixed effects models: A tutorial. *Journal of Cognition*, *1*, Article 9. https://doi.org/10.5334/joc.10

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*, 635–662. https://doi.org/10.1017/S0272263121000127

Buffington, J., & Morgan-Short, K. (2022). *Data and analysis: The reliability and validity of procedural memory assessments used in second language acquisition research* [Data set]. Open Science Framework. https://osf.io/ux4qs/

Burgoyne, A. P., Mashburn, C. A., Tsukahara, J. S., & Engle, R. W. (2022). Attention control and process overlap theory: Searching for cognitive processes underpinning the positive manifold. *Intelligence*, *91*, Article 101629. https://doi.org/10.1016/j.intell.2022.101629

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*, 685–709. https://doi.org/10.1007/s11336-013-9328-2

Clark, K., Birch-Hurst, K., Pennington, C. R., Petrie, A. C., Lee, J. T., & Hedge, C. (2022). Test-retest reliability for common tasks in vision science. *Journal of Vision*, *22*, Article 18. https://doi.org/10.1167/jov.22.8.18

Cohen, A. D., & Macaro, E. (2013). Research methods in second language acquisition. In E. Macaro (Ed.), *Continuum companion to second language acquisition* (pp. 107–136). Continuum.

Cronbach, L. J., & Furby, L. (1970). How we should measure "change": Or should we? *Psychological Bulletin*, *74*, 68–80. https://doi.org/10.1037/h0029382

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

DeBruine, L. M. (2020). *Simulating mixed effects.* https://debruine.github.io/tutorials/sim-lmer.html

DeBruine, L. M., & Barr, D. J. (2021). Understanding mixed-effects models through data simulation. *Advances in Methods and Practices in Psychological Science*, *4*, Article 251524592096511. https://doi.org/10.1177/2515245920965119

Draheim, C., Mashburn, C. A., Martin, J. D., & Engle, R. W. (2019). Reaction time in differential and developmental research: A review and commentary on the problems and alternatives. *Psychological Bulletin*, *145*, 508–535. https://doi.org/10.1037/bul0000192

Draheim, C., Tsukahara, J. S., & Engle, R. W. (2022). *Replication and extension of the toolbox approach to measuring attention control.* PsyArXiv. https://doi.org/10.31234/osf.io/gbnzh

Draheim, C., Tsukahara, J. S., Martin, J. D., Mashburn, C. A., & Engle, R. W. (2021). A toolbox approach to improving the measurement of attention control. *Journal of Experimental Psychology: General*, *150*, 242–275. https://doi.org/10.1037/xge0000783

Elgort, I. (2011). Deliberate learning and vocabulary acquisition in a second language. *Language Learning*, *61*, 367–413. https://doi.org/10.1111/j.1467-9922.2010.00613.x

Elgort, I., Brysbaert, M., Stevens, M., & Van Assche, E. (2018). Contextual word learning during reading in a second language: An eye-movement study. *Studies in Second Language Acquisition*, *40*, 341–366. https://doi.org/10.1017/S0272263117000109

Ellis, R. (2005). Measuring implicit and explicit knowledge of a second language: A psychometric study. *Studies in Second Language Acquisition*, *27*, 141–172. https://doi.org/10.1017/S0272263105050096

Fang, S., & Wu, Z. (2022a). *L2 predictive processing* [Data set]. https://osf.io/abhjv/

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

Faretta-Stutenberg, M., & Morgan-Short, K. (2018). The interplay of individual differences and context of learning in behavioral and neurocognitive second language development. *Second Language Research*, *34*, 67–101. https://doi.org/10.1177/0267658316684903

Godfroid, A. (2020). Sensitive measures of vocabulary knowledge and processing. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 433–453). Routledge. https://doi.org/10.4324/9780429291586-28

Godfroid, A., & Kim, K. M. (2021). The contributions of implicit-statistical learning aptitude to implicit second-language knowledge. *Studies in Second Language Acquisition*, *43*, 606–634. https://doi.org/10.1017/S0272263121000085

Granena, G. (2013). Individual differences in sequence learning ability and second language acquisition in early childhood and adulthood: Sequence learning ability and SLA. *Language Learning*, *63*, 665–703. https://doi.org/10.1111/lang.12018

Gulliksen, H. (1950). *Theory of mental tests*. Wiley.

Haines, N., Kvam, P. D., Irving, L. H., Smith, C., Beauchaine, T. P., Pitt, M. A., Ahn, W.-Y., & Turner, B. M. (2020). *Theoretically informed generative models can advance the psychological and brain sciences: Lessons from the reliability paradox*. PsyArXiv. https://doi.org/10.31234/osf.io/xr7y3

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

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*, 1448–1469. https://doi.org/10.1037/xlm0001028

Hox, J., Moerbeek, M., & van de Schoot, R. (2018). *Multilevel analysis techniques and applications*. Routledge.

Hui, B., Godfroid, A., & Elgort, I. (2022a). *A construct validation study of time-sensitive word measures*. Open Science Framework. https://doi.org/10.31219/osf.io/dwjmn

Hui, B., Godfroid, A., & Elgort, I. (2022b). *Materials, data, and analysis code for the paper: A construct validation study of time-sensitive word measures* [Data set]. Open Science Framework. https://osf.io/uyfh5/

Hui, B., & Huntley, E. (2021). Promoting open science practices: What can (should) graduate programs do? In L. Plonsky (Ed.), *Open science in applied linguistics*. John Benjamins.

Hui, B., Koh, J., & Ogawa, S. (2023). Voices of three junior scholars: A commentary on "(Why) are open research practices the future for the study of language learning?". *Language Learning*. Advance online publication. https://doi.org/10.1111/lang.12571

Hutchison, K. A., Balota, D. A., Cortese, M. J., & Watson, J. M. (2008). Predicting semantic priming at the item level. *Quarterly Journal of Experimental Psychology*, *61*, 1036–1066. https://doi.org/10.1080/17470210701438111

In'nami, Y., Mizumoto, A., Plonsky, L., & Koizumi, R. (2022). Promoting computationally reproducible research in applied linguistics: Recommended practices and considerations. *Research Methods in Applied Linguistics*, *1*, Article 100030. https://doi.org/10.1016/j.rmal.2022.100030

Isbell, D. R. (2021). *Open science, data analysis, and data sharing*. Open Science Framework. https://doi.org/10.31219/osf.io/pdj9y

Krus, D. J., & Helmstadter, G. C. (1993). The problem of negative reliabilities. *Educational and Psychological Measurement*, *53*, 643–650. https://doi.org/10.1177/0013164493053003005

Lammertink, I., Boersma, P., Wijnen, F., & Rispens, J. (2020). Statistical learning in the visuomotor domain and its relation to grammatical proficiency in children with and without developmental language disorder: A conceptual replication and meta-analysis. *Language Learning and Development*, *16*, 426–450. https://doi.org/10.1080/15475441.2020.1820340

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

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., & DeKeyser, R. M. (2020). Conflicting evidence of explicit and implicit knowledge from objective and subjective measures. *Studies in Second Language Acquisition*, *42*, 359–382. https://doi.org/10.1017/S0272263119000615

Mair, P., & Wilcox, R. (2020). Robust statistical methods in R using the WRS2 package. *Behavior Research Methods*, *52*, 464–488.

Marsden, E., Mackey, A., & Plonsky, L. (2016). The IRIS Repository: Advancing research practice and methodology. In A. Mackey & E. Marsden (Eds.), *Advancing methodology and practice: The IRIS repository of instruments for research into second languages* (pp. 1–21). Routledge.

Marsden, E. & Morgan-Short, K. (2023). (Why) are open research practices the future for the study of language learning?. *Language Learning*. Advance online publication. https://doi.org/10.1111/lang.12568

Marsden, E., Thompson, S., & Plonsky, L. (2018). A methodological synthesis of self-paced reading in second language research. *Applied Psycholinguistics*, *39*, 861–904. https://doi.org/10.1017/S0142716418000036

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

Miller, J., & Ulrich, R. (2013). Mental chronometry and individual differences: Modeling reliabilities and correlations of reaction time means and effect sizes. *Psychonomic Bulletin & Review*, *20*, 819–858. https://doi.org/10.3758/s13423-013-0404-5

Nash, J., & Varadhan, R. (2011). Unifying optimization algorithms to aid software system users: Optimx for R. *Journal of Statistical Software*, *43*, 1–14. https://doi.org/10.18637/jss.v043.i09

Nicklin, C., & Vitta, J. P. (2021). Effect-driven sample sizes in second language instructed vocabulary acquisition research. *The Modern Language Journal*, *105*, 218–236. https://doi.org/10.1111/modl.12692

Oliveira, C. M., Hayiou-Thomas, M. E., & Henderson, L. (2022). *Reliability of the serial reaction time task: If at first you don't succeed, try try try again.* PsyArXiv. https://doi.org/10.31234/osf.io/hqmy7

Paap, K. R., Johnson, H. A., & Sawi, O. (2016). Should the search for bilingual advantages in executive functioning continue. *Cortex*, *74*, 305–314. https://doi-org.proxy-um.researchport.umd.edu/10.1016/j.cortex.2015.09.010

Parsons, S., Kruijt, A. W., & Fox, E. (2019). Psychological science needs a standard practice of reporting the reliability of cognitive-behavioral measurements. *Advances in Methods and Practices in Psychological Science*, *2*, 378–395. https://doi.org/10.1177/2515245919879695

Patterson, A. S., & Nicklin, C. (2023). L2 self-paced reading data collection across three contexts: In-person, online, and crowdsourcing. *Research Methods in Applied Linguistics*, *2*, Article 100045. https://doi.org/10.1016/j.rmal.2023.100045

Plonsky, L., & Derrick, D. J. (2016). A meta-analysis of reliability coefficients in second language research. *The Modern Language Journal*, *100*, 538–553. https://doi.org/10.1111/modl.12335

Plonsky, L., Marsden, E., Crowther, D., Gass, S. M., & Spinner, P. (2020). A methodological synthesis and meta-analysis of judgment tasks in second language research. *Second Language Research*, *36*, 583–621. https://doi.org/10.1177/0267658319828413

R Core Team. (2022). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing. https://www.R-project.org/

Rouder, J. N., de la Pena, A. C., Pratte, M., Richards, V., Hernan, M., Pascoe, M., & Thapar, A. (2022). *Is the antisaccade task a unicorn task for measuring cognitive control?* Open Science Framework. https://doi.org/10.31219/osf.io/fhg3n

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

Siegelman, N., Bogaerts, L., & Frost, R. (2017). Measuring individual differences in statistical learning: Current pitfalls and possible solutions. *Behavioral Research*, *49*, 418–432. https://doi.org/10.3758/s13428-016-0719-z

Siyanova-Chanturia, A. (2013). Eye-tracking and ERPs in multi-word expression research: A state-of-the-art review of the method and findings. *The Mental Lexicon*, *8*, 245–268. https://doi.org/10.1075/ml.8.2.06siy

Siyanova-Chanturia, A., & Martinez, R. (2014). The idiom principle revisited. *Applied Linguistics*, *36*, 549–569. https://doi.org/10.1093/applin/amt054

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, *11*, 702–712. https://doi.org/10.1177/1745691616658637

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103*, 677–680. https://doi.org/10.1126/science.103.2684.677

Suzuki, Y. (2017). Validity of new measures of implicit knowledge: Distinguishing implicit knowledge from automatized explicit knowledge. *Applied Psycholinguistics*, *38*, 1229–1261. https://doi.org/10.1017/S014271641700011X

Suzuki, Y., Jeong, H., Cui, H., Okamoto, K., Kawashima, R., & Sugiura, M. (2022). An fMRI validation study of the word-monitoring task as a measure of implicit knowledge: Exploring the role of explicit and implicit aptitudes in behavioral and neural processing. *Studies in Second Language Acquisition*, *45*, 109–136. https://doi.org/10.1017/S0272263122000043

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

Trofimovich, P., & McDonough, K. (Eds.). (2011). *Applying priming methods to L2 learning, teaching and research: Insights from psycholinguistics*. John Benjamins.

Vafaee, P., Suzuki, Y., & Kachisnke, I. (2017). Validating grammaticality judgment tests: Evidence from two new psycholinguistic measures. *Studies in Second Language Acquisition*, *39*, 59–95. https://doi.org/10.1017/S0272263115000455

Verhaeghen, P., & De Meersman, L. (1998). Aging and the negative priming effect: A meta-analysis. *Psychology and Aging*, *13*, 435–444. https://doi.org/10.1037/0882-7974.13.3.435

Vitta, J. P., Nicklin, C., & McLean, S. (2021). Effect size-driven sample-size planning, randomization, and multisite use in L2: Instructed vocabulary acquisition experimental samples. *Studies in Second Language Acquisition*, *44*, 1424–1448. https://doi.org/10.1017/S0272263121000541

West, G., Shanks, D., & Hulme, C. (2018). Sustained attention, not procedural learning, is a predictor of reading, language and arithmetic skills in children. *Scientific Studies of Reading*, *25*, 47–63. https://doi.org/10.1080/10888438.2020.1750618

West, G., Vadillo, M. A., Shanks, D. R., & Hulme, C. (2017). The procedural learning deficit hypothesis of language learning disorders: We see some problems. *Developmental Science*, *21*, 1–13. https://doi.org/10.1111/desc.12552

Weigard, A., Clark, D. A., & Sripada, C. (2021). Cognitive efficiency beats top-down control as a reliable individual difference dimension relevant to self-control. *Cognition*, *215*, Article 104818. https://doi.org/10.1016/j.cognition.2021.104818

Winter, B. (2019). *Statistics for linguists: An introduction using R* (1st ed.). Routledge. https://doi.org/10.4324/9781315165547

Cite this article: Hui, B. and Wu, Z. (2023). Estimating reliability for response-time difference measures: Toward a standardized, model-based approach. *Studies in Second Language Acquisition,* 1–24. https://doi.org/10.1017/S027226312300027X