Annual Review of Applied Linguistics (2024), 1–11
doi:10.1017/S0267190524000047

CAMBRIDGE UNIVERSITY PRESS

RESEARCH ARTICLE

Reflecting on the use of response times to index linguistic knowledge in SLA

Bronson Hui and Ruirui Jia

University of Maryland, USA
Corresponding author: Bronson Hui; Email: bhui@umd.edu

Abstract

Response times (RTs) have become ubiquitous in second language acquisition (SLA) research, providing empirical evidence for the theorization of the language learning process. Recently, there have been discussions of some fundamental psychometric properties of RT data, including, but not limited to, their reliability and validity. In this light, we take a step back to reflect on the use of RT data to tap into linguistic knowledge in SLA. First, we offer a brief overview of how RT data are most commonly used as vocabulary and grammar measures. We then point out three key limitations of such uses, namely that (a) RT data can lack substantive importance without considerations of accuracy, (b) RT differences may or may not be a satisfactory psychometric individual difference measure, and (c) some tasks designed to elicit RT data may not be sufficiently fine-grained to target specific language processes. Our overarching goal is to enhance the awareness among SLA researchers of these issues when interpreting RT results and stimulate research endeavors that delve into the unique properties of RT data when used in our field.

Keywords: response time data; accuracy; second language acquisition

Response times (RTs) have been used widely in the field of second language acquisition (SLA), evidenced by book-length treatments on the subject (e.g., Jiang, 2013; Trofimovich & McDonough, 2011) as well as methodological introductions (e.g., Godfroid, 2020; Roberts, 2012). RT methods have a long history in neighboring fields such as psycholinguistics and bilingualism research to understand, for example, how more than one language is represented and organized in the mind, as well as the dynamics of the interaction between the languages (see Jiang, 2023 for an overview). For instance, research on the bilingual lexicon has shown that lexical access is nonselective; in other words, the language not in use (i.e., not being purposefully elicited or drawn on) within an experimental task is generally activated even when only one

language is relevant to the task at hand. Slower RTs to stimuli that are related to the non-target language (e.g., interlingual homophones and homographs or nonwords that are phonotactically legal in the not-in-use language) are taken as evidence that the nontarget language is activated. Findings such as these have served as the empirical basis for theoretical models of the bilingual lexicon, such as the *bilingual model of lexical access* (Grosjean, 1988, 1997, 2001) and the *bilingual interactive activation* (BIA) model (Grainger, 1993; Grainger & Dijkstra, 1992) and its later, expanded *BIA+* version (Dijkstra & van Heuven, 2002).

In SLA, RT measures have been adapted to index lexical and grammatical knowledge, complementing the use of untimed, paper-based tests centered on accuracy. This is partly because an important goal of SLA is to understand how learners acquire knowledge that can be deployed in authentic communication, characterized by a fair amount of time pressure. Therefore, RT measures present themselves as an attractive option because the tasks used to elicit RT data often do involve time pressure at the trial level (e.g., participants are often told to respond as quickly and accurately as possible), offering relatively more face validity than untimed tasks where the participants can take their time to reflect on their responses, which is almost inconceivable in dynamic language use. Furthermore, most of the psycholinguistic operations investigated by psycholinguists are outside of the participant's awareness, which coincides with the theorization of explicit and implicit knowledge in SLA (i.e., knowledge with and without awareness, respectively). Therefore, RT measures have been useful indirect tests that are believed to tap into implicit knowledge (Rebuschat, 2013). In summary, RT measures have been popular in SLA because they likely index the kind of lexical and grammatical knowledge that is readily available for relatively effortless, efficient processing and therefore authentic language use (e.g., Hui & Wu, 2024; Suzuki, 2017). Although there are various RT measures used in SLA, we summarize two broadly defined categories: primary and secondary RT measures.

Primary RT measures represent the use of simple reaction times, logged from the presentation of the stimulus to a button press, to directly index processing and/or retrieval speed. For example, researchers have used lexical decision tasks or timed yes/no vocabulary tests to tap into the quality of lexical knowledge as indicated by the speed of response (e.g., Hui & Godfroid, 2021; Pellicer-Sánchez & Schmitt, 2012; Suzuki & Kormos, 2023). To measure grammatical knowledge, several types of instruments have been employed, including grammaticality judgments, sentence continuation, and maze tasks – where participants choose one of the two presented options to grammatically continue a sentence (e.g., Suzuki & Kormos, 2023; Suzuki & Sunada, 2018). RT data in this context have been interpreted as the efficiency of processing morphosyntactic knowledge.

In terms of secondary measures, a computation from simple RTs is needed. One example is *RT differences*. Specifically, researchers model RT differences between trial types or experimental conditions. Depending on the experimental paradigm, faster or slower responses are expected. Using the priming paradigm, for instance, researchers ask participants to press the corresponding key to make a judgment on the target, such as whether a letter string forms a word (i.e., a lexical decision). Faster responses are expected when the target is preceded by a prime that is orthographically, phonologically, and/or semantically related to it. The idea is that the associations between the

https://doi.org/10.1017/S0267190524000047 Published online by Cambridge University Press

prime and the target drives a facilitation in processing, or priming, to be observed. Once observed, the priming effect can serve as evidence for knowledge of the prime and/or the target. More specifically, priming measures have been used, for instance, to indicate if a newly learned vocabulary item has been integrated into the learners' mental lexicon, or if there are robust the L2 form–meaning connections (e.g., Elgort, 2011; Hui et al., 2022). Another common experimental paradigm is the *violation paradigm*. The general idea is that learners' processing slows down when encountering ungrammaticality, provided that they have knowledge of the target structure. Commonly used tasks include self-paced reading and word-monitoring tasks. The former involves participants reading sentences divided into segments in a self-paced manner (i.e., using button presses), while the latter requires the participant to listen to sentences and monitor for a given target word to which they respond with a button press. In both cases, researchers expect a slowdown in processing the stimuli containing a grammatical error, compared with control trials without errors (e.g., Godfroid & Kim, 2021; Granena, 2013; Maie & DeKeyser, 2020). Again, this slowdown, when observed, is taken as evidence of knowledge of the target grammatical structure that is deployed in authentic language use (i.e., reading and listening).

Another example of secondary RT measures is the *coefficient of variation* (or CV; Segalowitz & Segalowitz, 1993), computed as the ratio of the standard deviation to the mean (or $CV = \frac{standard\ deviation}{mean}$) of a participant's RT distribution collected from judgment tasks. CV has been used as a processing stability measure in the sense that less variation in RTs represents more stable processing. Skill acquisition theory (DeKeyser, 2020) predicts a decrease in both simple RT and CV, as well as a positive correlation between them as learners develop proficiency (e.g., Hulstijn et al., 2009; Lim & Godfroid, 2015). This is because as learners automatize their knowledge, they are expected to process linguistic information in a faster and more stable manner, two key manifestations of automatic processing. Therefore, when there is a positive RT–CV correlation, CV is further interpreted as a processing automaticity measure in the sense that a lower CV value indicates more automatic processing at various linguistic levels (e.g., Hui, 2020; McManus & Marsden, 2019; Saito et al., 2023).

Given the widespread use of both primary and secondary RT measures to index linguistic knowledge, much theorization in SLA depends on the reliability and validity of RT data. Indeed, researchers are reminded that we should not take measures at their face value. On the contrary, constant scrutiny of reliability and validity should be carried out to ensure that the empirical base for theorization is strong (e.g., Cohen & Macaro, 2013; Hamrick, 2022). With some discussion on the reliability of RT measures already put forward (e.g., Hui & Wu, 2024), we focus more on the validity of RT measures in this paper.

RT data processing and accuracy

First and foremost, RT data need to be considered in conjunction with accuracy data. That is, RT data in isolation can lack substantive significance. A key reason is that, in many, if not all, cases, researchers preprocess RT data such that only correct responses are analyzed. This step is potentially meant to align the data preprocessing procedure with conventional practices in psycholinguistics where incorrect responses are

https://doi.org/10.1017/S0267190524000047 Published online by Cambridge University Press

considered to be an indication that the participant engaged in nontarget psycholinguistic operations, and hence are irrelevant to the phenomenon in question. However, in SLA, researchers are most interested in learning *gains* and, as a result, SLA studies often involve some form of progression from no knowledge to partial knowledge to mastery, as typically demonstrated by an improvement in accuracy. In other words, a high accuracy rate is not guaranteed, which is in sharp contrast with psycholinguistic research, where a relatively high accuracy is almost a prerequisite for an effect. A key implication of discarding incorrect trials is that a lower accuracy means a reduced number of observations for analysis, not only lowering statistical power but also potentially shifting the inference made by the researchers. In this light, interpretations of RT data in the same way as psycholinguists require some caution.

For example, if we were to take RT data as a speed measure at its face value, one might suggest that two groups of learners with a similar mean RT are at a comparable level of mastery, at least in terms of processing speed. However, one group may be more accurate in its responses, while the other is less accurate. When that is the case, believing that they are indistinguishable in general terms can be problematic. The seeming equivalence in performance is merely an artifact of the data preprocessing, namely removing incorrect responses from the analysis. The RT data for the low-accuracy group lose representativeness disproportionately as more data are discarded for them. On this account, using processing speed (alone) to predict language use performance requires more caution, too (e.g., Hui & Godfroid, 2021; Zhang & Yang, 2023), because a fast responder might not be a more proficient learner, as hypothesized. They can be merely an individual who compromises accuracy for speed. Indeed, this accuracy–speed trade-off is not new to researchers using RT data, but it has not received sufficient attention, in part, because the modeling approach undertaken can almost always handle only one outcome variable. As a result, variability due to leaning toward accuracy or speed is somewhat accepted as measurement error in practice.

To further illustrate the relationship between RT and accuracy, we plot the open data shared by Hui (2020), who investigated the processing trajectory of intentional word learning (see Figure 1). In the study, the author trained English speakers to learn 16 Swahili–English translation pairs, after which the participants performed 10 blocks of *animacy judgments* (i.e., to decide whether the stimulus denotes a living thing). Overall, the accuracy gradually increased as the learners took advantage of the feedback given after each trial. RT of correct responses decreased, more obviously during the first blocks. From these patterns, suggesting that there was "faster processing by the participants over time" is not wrong (Hui, 2020, p. 340). However, the more accurate and complete description would be that processing of the words the participants have learned *and responded correctly to* was faster over time. The key here is that researchers must match their interpretations of the results with their data analysis approach, including the preprocessing steps. It is because when incorrect trials are excluded, researchers artificially create systematic missingness in the data, changing the inference from what is initially intended (i.e., from processing speed for all items to that only for correctly responded items).

This need to consider accuracy along with RT also applies to secondary measures, such as lexical decisions in the priming paradigm. As discussed, when priming is observed, it means that the lexical item is represented in the mental lexicon and is

https://doi.org/10.1017/S0267190524000047 Published online by Cambridge University Press

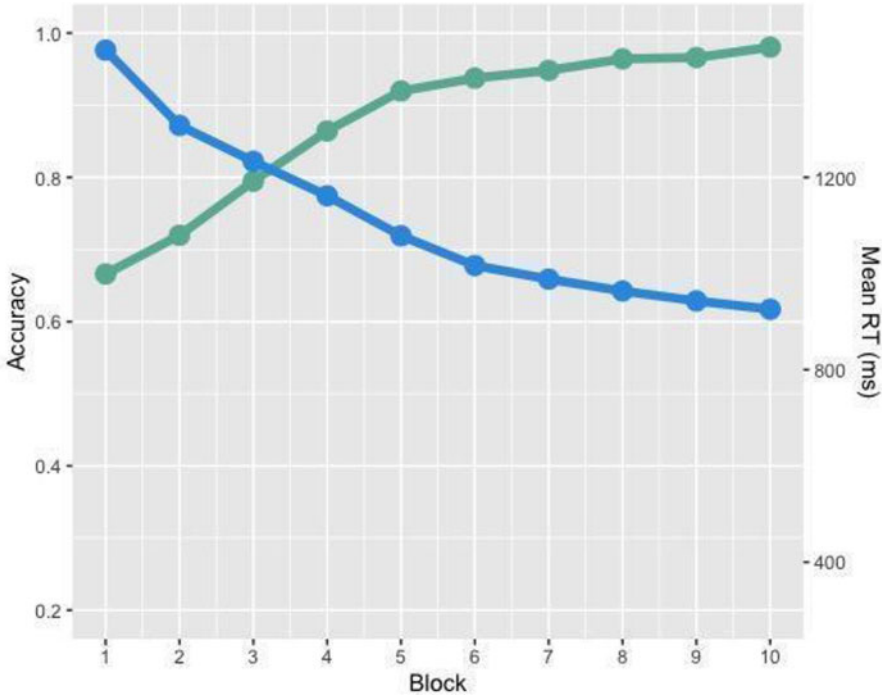

Figure 1. Accuracy (green line) and RT (blue line) trajectories in intentional word learning reported in Hui (2020).

connected to other items in the mind, depending on the manipulation. In this regard, priming has been taken as evidence of lexical knowledge. Perhaps more importantly, priming measures have been interpreted as tapping into tacit or nondeclarative lexical knowledge, especially when the prime is masked by symbols and a very brief presentation, such that the participant is not consciously aware of its presence. It is because, first, the psycholinguistic operations that drive the facilitation are out of the participant's awareness (e.g., Elgort, 2011; Hui et al., 2022). Second, the participants are almost always unaware of the masked prime (or, at least, the researcher would confirm that they are) and still show a priming effect. At the same time, it is useful to bear in mind that this interpretation applies only to the items that the participants successfully recognize in the first place. This is because, again, only correct responses are included in the analysis.

Crucially, some theoretical tensions start to emerge when considering the above. On the one hand, a lexical decision task itself is an explicit task. The participants are well aware of their judgments. Correct responses then require explicit knowledge. On the other hand, tacit knowledge is believed to be categorically different from explicit knowledge although they are intimately related (Hui et al., 2022). But the measurement of it (lexical decisions in the priming paradigm) requires the participant to first demonstrate explicit knowledge through a correct response. In a sense, it seems that

https://doi.org/10.1017/S0267190524000047 Published online by Cambridge University Press

researchers may operate under the assumption that the development of tacit knowledge first requires explicit knowledge. Whether this assumption holds is an ongoing debate in the literature. What we want to stress here is that researchers using the priming paradigm to measure tacit or implicit knowledge must be aware of this assumption, which they inherently accept through their data analysis procedure.

Finally, a closer look at Figure 1 suggests that the decrease in RT of correct responses was relatively more obvious during the initial stages and gradually leveled off. This was also the stage in which the accuracy was relatively low in the trajectory. In other words, the most noticeable result (i.e., development of speed) was observed during a stage where RT data were less informative due to low accuracy, creating an interpretative RT–accuracy dilemma for researchers using RT data to index learning. An important implication of the need to consider accuracy and RT in tandem is that RT measures might be more suitable for examining the refinement of *partial knowledge* in the context of a word learning study. If researchers are interested in earlier stages of acquisition where accuracy can be low, they should consider measures that combine speed and accuracy such that, for example, processing speed is adjusted for error rates (e.g., $\frac{mean\ RT}{error\ rate}$) (see Vandierendonck, 2017 for an overview).

RT differences as an individual measure

Another issue to consider is the extent to which RT differences can be used as an individual difference measure, that is, a measure that can index a given learner's ability or knowledge levels. In almost all areas of SLA, a high correlation is expected between the learner's test score and their level of knowledge or skills. Someone achieving higher on a vocabulary test, for example, should possess more or better lexical knowledge than someone who has a lower score. In this regard, SLA researchers must maintain a high measurement bar for our measures. Measurement, by definition, is the principled assignment of numerical values to attributes, such as L2 knowledge (Stevens, 1946). Unfortunately, this basic principle of measurement does not necessarily apply to RT differences, at least in the ways that they are used in SLA. Researchers, for instance, cannot necessarily claim that a learner who slows down more when encountering ungrammaticality in a self-paced reading task has more knowledge than someone who slows down less. Similarly, a larger priming effect with lexical decisions (more facilitation) may not be interpreted as the learner having better lexical knowledge. This is especially true when RT measures are more sensitive to factors such as handedness, physical difficulties, and coordination.

Part of the issue is that some of the paradigms SLA researchers use to capture RT differences have been borrowed from psycholinguistics, where researchers primarily focus on group-level effects. A higher degree of unreliability is accepted because the performance of individuals is influenced by factors that the researcher cannot control, such as the familiarity of the stimuli by the participant, the order of presentation, and so on. Also, psycholinguists often adopt a counterbalanced design in which participants only see one version of an item to avoid multiple exposures. As a result, researchers understand their limitations when it comes to examining the RT difference of an individual. Some would even argue that RT differences are only meant to be analyzed at the group level, although others still use them to index individual ability in a moderation

https://doi.org/10.1017/S0267190524000047 Published online by Cambridge University Press

analysis to investigate differential effects of factors, such as proficiency, on priming, for example. In most cases, psycholinguists also seldom use RT differences as a predictor variable, which assumes perfect reliability.

In this light, SLA researchers must decide for ourselves how we want to use RT differences and understand the consequences of accepting RT differences as merely a group-level measure. First of all, to be able to examine aptitude-treatment interactions (e.g., whether a given treatment is particularly useful for someone with certain characteristics; see DeKeyser, 2021), researchers must have an outcome measure that can index an individual's ability. On the contrary, when RT differences are used as a group-level outcome, the researcher's hands are tied, because the moderation analysis can no longer reveal differential treatment effects that are based on the characteristics of the learner, such as perceptual abilities (e.g., sound discrimination skills; Shao et al., 2023) and cognitive abilities (e.g., working memory capacity; Suzuki & DeKeyser, 2017). Therefore, when the outcome does not tap into one's ability at the individual level, the analysis itself is not meaningful. This is an immensely important limitation, especially for researchers who rely on RT differences to tap into tacit lexical knowledge (with a lexical decision task in the priming paradigm), or implicit or automatized explicit grammatical knowledge (with a word-monitoring or a self-paced reading task). What this means is that the field can face challenges when conducting individual differences research in terms of how this type of knowledge is acquired, as measured by RT differences.

The second consequence of accepting RT differences as group-level measurse is that researchers cannot examine their predictive validity. The predictive validity of measures is critical because when a measure indexing a given construct can strongly predict overall language performance, researchers have empirical evidence that this construct is germane to authentic language use, an important goal of SLA research. For example, it is useful to know how relevant implicit grammatical knowledge is to reading and listening comprehension. In analytic terms, one would use the RT difference, indexing implicit knowledge, to predict comprehension performance in a regression model. But this analysis would require the RT measure to be perfectly reliable, a psychometric property that RT differences often lack and the very reason why psycholinguists only use RT differences at the group level. Indeed, many arguments for measuring tacit, implicit, or automatized explicit knowledge, using RT differences, are exactly that these knowledge types are key for real-time language use. However, when group-level RT differences cannot be scrutinized in terms of predictive validity (or be used to predict language use in a statisical model), such arguments can only remain at the conceptual level, which is truly undesirable.

Relatedly, if RT differences only represent a group-level measure, researchers are not in a position to test their construct validity – specifically, their convergence and divergence validities – in relation to other measures because the statistical technique often used to investigate construct validity, confirmatory factor analysis, relies on covariance between variables that index one's ability at the individual level. A group-level priming measure, for example, may not offer a starting point for examining its correlation with other tests. When that is the case, it is almost impossible to empirically differentiate these RT difference measures, supposedly tapping into tacit or implicit knowledge, from other assessments of explicit knowledge.

https://doi.org/10.1017/S0267190524000047 Published online by Cambridge University Press

In summary, SLA researchers must further understand RT difference measures and the price for accepting them as merely group-level measures. Collectively, we need to make informed decisions on how to use RT differences, or whether to use them at all. The first step in this direction is to properly examine the psychometric properties of RT differences. For example, there has already been work on estimating the reliability of RT differences using a model-based approach (Hui & Wu, 2024). If RT differences can be reliable, the next step is to more closely scrutinize their validity by, for example, correlating priming measures with existing vocabulary measures or by using RT differences to predict language performance. It is worth mentioning that the motivation of such work can be initially questioned, because those who assert that RT difference measures are designed for group-level analyses would consider this kind of investigation as a misuse of the data. At the same time, we stress that the stakes are high for SLA researchers to understand what RT differences are measuring at an individual level.

RT data involving multiple processes

Finally, reaction time can be less informative if a task involves more than one process (Jiang, 2022). When multiple processes are involved in a task, it is difficult to disentangle the observed effects to make meaningful interpretations. One example is a *self-paced reading task* (see Marsden et al., 2018 for an overview). As mentioned, self-paced reading tasks are often used to measure implicit grammatical knowledge deployed in real time. The idea is that learners with the target knowledge are expected to show sensitivity to grammatical violations, operationalized as a slowdown in reading times, compared with the baseline grammatical condition. Another example also involving sentence reading is a *grammaticality judgment task*, which has been used to measure explicit and implicit knowledge, depending on modality and time–pressure manipulations. In both self-paced reading and grammaticality judgment tasks, it can be difficult to pinpoint specifically what aspect of the stimuli, or what cognitive processes, lead to the delay in processing ungrammatical sentences, as sentence reading involves processes such as word recognition, parsing, and semantic integration. Therefore, it is perhaps not straightforward to take the tasks at face value and claim that these tasks afford relatively pure measures of grammatical knowledge.

A similar issue can be observed in vocabulary research. For example, the *word association task* is sometimes used to obtain RT data to index the strength of networks in a learner's mental lexicon. The key here is that researchers find it difficult to specify the factors and their contribution to the observed effects. For example, when a group of learners is faster, a number of interpretations are sensible, including more efficient word recognition, stronger associations, and/or a larger number of associations available. It is also challenging to spell out the specific level(s) of association (lexical vs. conceptual) the task is targeting. It must be acknowledged that these issues also apply to accuracy-based measures. These are perhaps not problems unique to RT research. The key message here is that to ensure validity of their measures (i.e., to make accurate interpretations of their results), SLA researchers should be mindful of the psychological processes involved in completing the tasks. While no measure is a pure measure of anything, knowing what is or can be underpinning a numerical result that we interpret is of paramount importance.

https://doi.org/10.1017/S0267190524000047 Published online by Cambridge University Press

Conclusion

In this paper, we started by pointing out that RT data play a critical role in the formation of an empirical base for the theorization of the language learning process. It is therefore important to examine and reflect upon the use of RT data, mostly elicited by tasks borrowed from our neighboring field of psycholinguistics. We presented three validity issues that deserve more attention: (a) RT data need to be considered along with accuracy data, (b) there is currently little research on the extent to which RT differences can measure an individual's ability, and (c) RT data elicited from tasks that involve multiple cognitive processes can be difficult to interpret. We hope to remind SLA researchers that while RT measures offer valuable insights, they should be approached with diligent attention to these three issues. SLA researchers must recognize the limitations and nuances that may not be relevant in the field from which we borrow these measures but are of paramount importance to ours. Future research should focus on scrutinizing psychometric properties, exploring alternative measures for different stages of acquisition, and enhancing the theoretical grounding of task selection in order to advance our understanding of language processing in SLA. We would like to close on a personal note; we have cited some previous work by the first author of this article, not (just) because he wants to receive one more citation count, but indeed to demonstrate that he, too, is guilty of perhaps not thinking about these issues carefully enough. As a field, we must collectively pay sufficient attention to the measures that we use. Every SLA researcher should be a methodologist in some way to move us all forward.

Acknowledgements. The authors would like to express our sincere gratitude to Dr. Nan Jiang for his input in conceptualizing this piece and to Drs Phil Hamrick and Yuichi Suzuki for their invaluable feedback on an earlier draft of this manuscript. All errors are ours.

References

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.

DeKeyser, R. (2020). Skill acquisition theory. In B. Van Patten, G. D. Keating & S. Wulff (Eds.), *Theories in second language acquisition: An introduction* (3rd ed., pp. 83–104). Routledge.

DeKeyser, R., (Ed.) (2021). *Aptitude-treatment interaction in second language learning*. John Benjamins Publishing Company.

Dijkstra, T., & van Heuven, W. J. B. (2002). The architecture of the bilingual word recognition system: From identification to decision. *Bilingualism: Language and Cognition*, *5*(3), 175–197.

Elgort, I. (2011). Deliberate learning and vocabulary acquisition in a second language. *Language Learning*, *61*(2), 367–413.

Godfroid, A. (2020). Sensitive measures of vocabulary knowledge and processing. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (1st ed., pp. 433–453). Routledge.

Godfroid, A., & Kim, K. M. (2021). The contributions of implicit-statistical learning aptitude to implicit second-language knowledge. *Studies in Second Language Acquisition*, *43*(3), 606–634.

Grainger, J. (1993). Visual word recognition in bilinguals. In R. Schreuder & B. Weltens (Eds.), *The bilingual lexicon* (pp. 11–25). John Benjamins.

Grainger, J., & Dijkstra, T. (1992). On the representation and use of language information in bilinguals. In R. J. Harris (Ed.), *Cognitive processing in bilinguals* (pp. 207–220). Elsevier.

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*(4), 665–703.

Grosjean, F. (1988). Exploring the recognition of guest words in bilingual speech. *Language and Cognitive Processes*, *3*(3), 233–274.

Grosjean, F. (1997). Processing mixed language: Issues, findings, and models. In A. M. B. de Groot & J. F. Kroll (Eds.), *Tutorials in bilingualism: Psycholinguistic perspectives* (pp. 225–254). Lawrence Erlbaum Associates.

Grosjean, F. (2001). The bilingual's language modes. In N. Janet (Ed.), *One language two languages* (pp. 1–22). Blackwell.

Hamrick, P. (2022). Conducting reaction time research in second language psycholinguistics. In A. Godfroid & H. Hopp (Eds.), *The Routledge handbook of SLA and psycholinguistics* (pp. 150–163). Routledge.

Hui, B. (2020). Processing variability in intentional and incidental word learning: An extension of Solovyeva and Dekeyser 2018. *Studies in Second Language Acquisition*, *42*(2), 327–357.

Hui, B., & Godfroid, A. (2021). Testing the role of processing speed and automaticity in second language listening. *Applied Psycholinguistics*, *42*(5), 1089–1115.

Hui, B., Godfroid, A., & Elgort, I. (2022). *A construct validation study of time-sensitive word measures* [Preprint]. Open Science Framework.

Hui, B., & Wu, Z. (2024). Estimating reliability for response-time difference measures: Towards a standardized, model-based approach. *Studies in Second Language Acquisition, 45*(1), 227–250.

Hulstijn, J. H., Van Gelderen, A., & Schoonen, R. (2009). Automatization in second language acquisition: What does the coefficient of variation tell us? *Applied Psycholinguistics*, *30*(4), 555–582.

Jiang, N. (2013). *Conducting reaction time research in second language studies* (1st ed.). Routledge.

Jiang, N. (2022). Synthesis: Research methods in L2 psycholinguistics. In A. Godfroid & H. Hopp (Eds.), *The Routledge handbook of SLA and psycholinguistics* (pp. 178–188). Routledge.

Jiang, N. (2023). *The study of bilingual language processing*. Oxford University Press.

Lim, H., & Godfroid, A. (2015). Automatization in second language sentence processing: A partial, conceptual replication of Hulstijn, Van Gelderen, and Schoonen's 2009 study. *Applied Psycholinguistics*, *36*(5), 1247–1282.

Maie, R., & DeKeyser, R. M. (2020). Conflicting evidence of explicit and implicit knowledge from objective and subjective measures. *Studies in Second Language Acquisition*, *42*(2), 359–382.

Marsden, E., Thompson, S., & Plonsky, L. (2018). A methodological synthesis of self-paced reading in second language research. *Applied Psycholinguistics*, *39*(5), 861–904.

McManus, K., & Marsden, E. (2019). Signatures of automaticity during practice: Explicit instruction about L1 processing routines can improve L2 grammatical processing. *Applied Psycholinguistics*, *40*(1), 205–234.

Pellicer-Sánchez, A., & Schmitt, N. (2012). Scoring Yes–No vocabulary tests: Reaction time vs. nonword approaches. *Language Testing*, *29*(4), 489–509.

Rebuschat, P. (2013). Measuring implicit and explicit knowledge in second language research. *Language Learning*, *63*(3), 595–626.

Roberts, L. (2012). Psycholinguistic techniques and resources in second language acquisition research. *Second Language Research*, *28*(1), 113–127.

Saito, K., Uchihara, T., Takizawa, K., & Suzukida, Y. (2023). Individual differences in L2 listening proficiency revisited: Roles of form, meaning, and use aspects of phonological vocabulary knowledge. *Studies in Second Language Acquisition*, 1–27, Advance online publication.

Segalowitz, N. S., & Segalowitz, S. J. (1993). Skilled performance, practice, and the differentiation of speed-up from automatization effects: Evidence from second language word recognition. *Applied Psycholinguistics*, *14*(3), 369–385.

Shao, Y., Saito, K., & Tierney, A. (2023). How does having a good ear promote instructed second language pronunciation development? Roles of domain-general auditory processing in choral repetition training. *TESOL Quarterly*, *57*(1), 33–63.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103*(2684), 677–680.

Suzuki, Y. (2017). Validity of new measures of implicit knowledge: Distinguishing implicit knowledge from automatized explicit knowledge. *Applied Psycholinguistics*, *38*(5), 1229–1261.

Suzuki, Y., & DeKeyser, R. (2017). Exploratory research on second language practice distribution: An Aptitude × Treatment interaction. *Applied Psycholinguistics*, *38*(1), 27–56.

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

Suzuki, Y., & Sunada, M. (2018). Automatization in second language sentence processing: Relationship between elicited imitation and maze tasks. *Bilingualism: Language and Cognition*, *21*(1), 32–46.

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

Vandierendonck, A. (2017). A comparison of methods to combine speed and accuracy measures of performance: A rejoinder on the binning procedure. *Behavior Research Methods*, *49*(2), 653–673.

Zhang, Z., & Yang, X. (2023). Using sentence processing speed and automaticity to predict L2 performance in the productive and receptive tasks. *International Review of Applied Linguistics in Language Teaching*, 1–27, Advance online publication.

Cite this article: Hui, B and Jia, R. (2024). Reflecting on the use of response times to index linguistic knowledge in SLA. *Annual Review of Applied Linguistics*, 1–11. https://doi.org/10.1017/S0267190524000047