Reading in a Foreign Language March 2024, Volume 36, No. 1
ISSN 1539-0578 pp. 1–7

The Authors Respond: Issues Surrounding Reliability, Quality, and Practicality with Timed-Reading Assessments: Expanding on Carter et al.'s (2023) *A Unitary Measure of L2 Silent Reading Fluency Accounting for Comprehension*

Steven J. Carter
Brigham Young University-Hawaii
United States

Matthew P. Wilcox
Brigham Young University
United States

Abstract

Carter et al. (2023) presented empirical evidence in support of a proposed new measure of L2 silent reading fluency. Referencing their method, this article addresses three separate practical issues related to using timed readings (TRs) to foster L2 reading fluency: TR assessment reliability, quality, and practicality. One seeming limitation of Carter et al.'s (2023) method was the relatively low reliability of three separate TR quizzes used in their study on reading fluency. However, considering that the interpretation and use of reliability estimates should be context-dependent, we argue that the standard expectations of 0.8 or higher may be simply unrealistic given the unique constraints surrounding timed readings. Furthermore, reliability is only one facet of a validity argument and intentional changes aimed at increasing reliability may, at times, come at the expense of other important aspects of validity. This article also offers practical advice for constructing effective TR quiz questions and directs the reader to tools for tracking student readers' reading fluency progress.

Keywords: timed-reading assessment, silent reading fluency, assessment reliability, assessment validity, second language reading

In the October, 2023 issue of *Reading in a Foreign Language* Carter et al. proposed a new formula for measuring silent L2 reading fluency performance and progress and gave some initial empirical evidence of its validity when used with timed readings (TRs). The primary benefit of the new *rf* formula is that both rate and comprehension factor into the calculation. In essence it controls for comprehension in the measurement of reading fluency. The structure of the formula prevents high performance in one sub-skill (rate or comprehension) from compensating for low performance in the other. This is important because logic dictates that the silent reading fluency construct presupposes some level of intelligent processing of the text. In other words, it makes little sense to speak of reading rate without considering how much comprehension occurred. On the other hand, very slow reading with high comprehension cannot be described as fluent either.

The new formula could have value for both practitioners seeking to track readers' incremental progress, and for researchers interested in a measure of fluency potentially more accurate than rate alone.

What follows is a series of thoughtful practitioner-oriented questions and their respective answers, all pertinent to the Carter et al. (2023) study. We, the authors, appreciated the opportunity to respond to the questions. It allowed us to expand on important aspects of practical TR use in an applied setting and to focus on key points that could not be reasonably addressed in the article.

Question 1: You made reasonable efforts to revise your TR multiple-choice comprehension questions based on weaknesses you knew you had with your existing materials, yet you were still plagued with lower reliability than you wished to have. How would you revise your tests given what you know now?

Following data collection for Carter et al. (2023), we made substantive efforts to revise the TR quiz items used in the study and others not used in the study. However, despite these efforts, when we have administered the revised quizzes to populations similar to that in the study, we have only succeeded in reaching reliability scores as high as 0.60 or 0.70 for any single set of 10 quiz questions (and reaching 0.70 was a rare occurrence). To respond directly to the question, perhaps further revision would not be a productive course of action. Instead, it may be that revisiting the accepted standard for reliability is needed for this type of assessment. Our experience leads us to believe that perhaps, given the specific constraints surrounding the population, context, and task[1], practicality would dictate that lower reliability would be acceptable because it is incredibly difficult to achieve the typical standard for reliability within the associated constraints.

The common notion that an assessment must have a reliability coefficient of 0.80 or higher may be a good rule of thumb, but to apply it universally is shortsighted because the interpretation and use of reliability estimates is context-dependent (Cho & Kim, 2015). For example, Nunnally (1967) in his seminal work *Psychometric Theory* originally posited that lower reliability estimates, as low as 0.50, were adequate for exploratory research. Although he raised this level to 0.70 in his second and third editions, other more recent publications have repeated the claim that for exploratory research or other similar low-stakes situations, values as low as 0.60 are suitable (George & Mallery, 2003; Hair, 2010). Taber (2018) further shows that in a review of the literature that reliability estimates as low as 0.45 have been considered satisfactory in published research for certain situations. Conversely, the same authors cited above also posit that for contexts such as applied research or high-stakes testing, reliability estimates much higher than 0.80 are needed. In all of this, researchers should determine acceptable reliability value ranges prior to data collection and analysis, based on prior research, the purpose(s) of the study, and the stakes, decisions, and consequences of the results (Cortina, 1993). For this study, we hoped for higher reliability values, but were willing to accept lower values knowing that this study was exploratory.

There are, however, other considerations that are relevant, specifically concerning reliability as a facet of validity. While there are multiple definitions, frameworks, and approaches, validity can be defined as a set of activities or investigations involving "the collection of different types of evidence in order to make a holistic evaluation of an assessment's fitness for a purpose" (Schmidgall & Xi, 2022, p. 1). Reliability, therefore, should be one of many pieces of evidence for an assessment's validity, and a high reliability coefficient does not always necessarily reflect the quality or relevance of test content. To use a common saying, there are ways of gaming the system to achieve high reliability. An assessment could consistently measure an aspect of language ability that is not relevant or important for the intended purpose of the assessment (Im et al., 2019) or other qualitative aspects such as test-taker engagement, test fairness, and the impact of cultural and linguistic diversity on test performance (International Test Commission [ITC], 2019; Swain, 1993). For example, we have observed test developers who have artificially increased reliability estimates by simply using the "alpha-if-deleted" function provided by many statistical packages, with little or no consideration for the content.

Further, some reliability estimates can be inflated by simply adding more items, again with little regard to their quality (Kopalle & Lehmann, 1997; Schmitt, 1996; Tavakol & Dennick, 2011). While we could have intentionally included a class of very low-ability readers in the study or added more items to the quizzes to increase reliability, there were practical and ethical considerations and constraints that logically contradicted those courses of action. First, the TRs would not be appropriate for very low-ability readers. As such, including these readers in the study would have been ethically questionable. Second, there is a reasonable limit to the number of items that can be developed from a single 1000-word reading without violating local or item independence (Ha, 2022). Third, expecting readers to retain reading material in their memory sufficiently well to answer more than ten questions without referring to the text is likely an unrealistically high expectation.

All things considered, including the necessity of keeping TRs and their associated quizzes practical, it may be reasonable to think of a lower threshold of reliability as satisfactory for TRs.

Question 2: Since readers/language teachers/testing practitioners may need to develop, adapt, or adopt reading passages and reading comprehension questions for their own procedures to use with your formula, what specific suggestions might you have for developing test questions based on the results of your study?

In our experience reviewing TR materials, many times associated TR quizzes have not been vetted and appear to be too easy. Items that seem too easy may be less of a concern with low-ability readers because they may discriminate—separate learners—reasonably well regardless (reliability hinges on adequate discrimination [Carr, 2011; Ebel, 1967]). However, for classes of mid- to high-ability readers, items that seem too easy are likely problematic due to poor discrimination.

Constructing well-functioning TR quiz items is perhaps more challenging than one might presume, especially for higher-ability readers from our experience. Among constraints already mentioned, items should meet the assumption of item independence (Carr, 2011). Second,

questions should not probe into obscure details that are unlikely to be noticed by readers. As was mentioned in Carter et al. (2023):

Initial reading purpose moderates comprehension by guiding what readers attend to within a text (De Hoyos & David, 2018; Erten, 2018; Grabe, 2009), but when the purpose is to practice or demonstrate fluency, readers are not primed to attend to specific details and it is unsurprising for them to recall content imperfectly. (p. 109).

Because readers cannot refer back to the reading when answering TR quiz items, the items should refer to more salient information that is more likely be noticed and stored in working memory rather than small details that may go unnoticed (Quinn et al., 2007). Ultimately, items must strike a delicate balance between not being too easy (i.e., requiring meaningful comprehension of the reading) and not being overly difficult either.

These constraints together, make the development of well-functioning TR quiz items difficult. However, we recommend a useful approach that we followed to identifying more salient or memorable information from which to write items:

1) First, we asked several colleagues to do a pilot reading of a given TR text. They were instructed to do it at a comfortable pace, without belaboring the reading. They read the text only one time.

2) Second, immediately after finishing the reading, we asked them to write down everything they remembered from the reading, including as many details as they could. They were encouraged to write without being concerned about punctuation, grammar, or neatness. Another possibility is to have pilot readers verbally record what they remember; however, we did not attempt this.

3) Third, after our pilot readers had finished writing their notes, we then compiled the notes of multiple pilot readers together and looked for common ideas and details in their notes on which to base questions.

It is also important to clearly identify the reading constructs being targeted by items and how those constructs will be operationalized in specifications (e.g., how many questions will target which construct) in order to ensure construct relevance (American Educational Research Association et al., 2014). Due to the need for item independence, we found it was difficult to develop more than two items per reading which focused on global understanding (i.e., main ideas). We found the same to be true for items focused on basic inferences. In essence, this necessitated that the remainder of the items be focused on details which needed to be salient for the reasons articulated above.

Ultimately, developing reasonably effective TR quizzes required a fair amount of trial and error. We would generally recommend some meaningful piloting of select TR quizzes especially if they are to be used for measurement and even more so if those measurements have meaningful effects on grades or placement.

Question 3: For practitioners who wish to use your formula, what practical advice can you offer? For instance, can you offer sample spreadsheets where sample data are input into the formula, and where data are plotted onto Anderson's fluency/comprehension quadrant?

We generally, encourage practitioners to make use of the Modified Reading Fluency Chart available in Appendix E of Carter et al. (2023). We would also encourage them to make use of the Reading Fluency Tracking Workbook (Carter & Anderson, 2021) or the simplified version of it; both are available at https://tinyurl.com/rdngfluencytools. Both of them map out results nicely in Excel, requiring little of the user other than inputting information. They both include links to short tutorial videos which explain briefly how to use them, and there are examples with information already inputted.

We would also encourage users not to overthink the use of the formula. Ours was a fairly complex approach to producing evidence of the validity of the formula. However, the end goal was to reach a solution that could be used with some straightforward confidence and simplicity.

Note

1.We are referring specifically to the context and task grouped with the higher-ability population in question. Whereas low-ability readers may select wrong answers because of very limited vocabulary and a general struggle with sentence-level comprehension, mid- to high- ability readers are less likely to do so. Consequently, writing well-functioning items that satisfy constraints for mid- to high-ability readers seems to require much more nuance and care. The result is that it is more difficult to produce items that discriminate well.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Bachman, L. F., & Palmer, A. S. (2009). *Language testing in practice: Designing and developing useful language tests*. Oxford University Press.

Carr, N. T. (2011). *Designing and analyzing language tests.* Oxford University Press.

Carter, S. J., & Anderson, N. J. (2021, March 24–27). *Equitable measurement of ELLs' reading fluency performance and progress* [Conference workshop]. TESOL 2021: International Convention & English Language Expo Conference, virtual conference.

Carter, S. J., Wilcox, M. P., & Anderson, N. J. (2023). A unitary measure of L2 silent reading fluency accounting for comprehension. *Reading in a Foreign Language, 35*(2), 106–137. https://hdl.handle.net/10125/67444

Cho, E., & Kim, S. (2015). Cronbach's coefficient alpha: Well known but poorly understood. *Organizational Research Methods*, *18*(2), 207–230. https://doi.org/10.1177/1094428114555994

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, *78*(1), 98–104. https://doi.org/10.1037/0021-9010.78.1.98

De Hoyos, D., & David, N. (2018). Understanding reading purpose. In J.I. Liontas, TESOL International Association, & M. DelliCarpini (Eds.), *The TESOL encyclopedia of English language teaching*. John Wiley & Sons, Inc. https://doi-org.byuh.idm.oclc.org/10.1002/9781118784235.eelt0501

Ebel, R. L. (1967). The relation of item discrimination to test reliability. *Journal of Educational Measurement, 4*(3), 125–128. https://www.jstor.org/stable/1434085

Erten, I.E. (2018). Activation of prior knowledge. In J.I. Liontas, TESOL International Association, & M. DelliCarpini (Eds.), *The TESOL encyclopedia of English language teaching*. John Wiley & Sons, Inc. https://doi-org.byuh.idm.oclc.org/10.1002/9781118784235.eelt0801

George, D., & Mallery, P. (2003). *SPSS for Windows step by step: A simple guide and reference, 11.0 update* (4th ed). Allyn and Bacon.

Grabe, W. (2009). *Reading in a second language: Moving from theory to practice.* Cambridge University Press.

Ha, H. T. (2022). Test format and local dependence of items revisited: A case of two vocabulary levels tests. *Frontiers in Psychology*, *12*, 1–6. https://doi.org/10.3389/fpsyg.2021.805450

Hair, J. F. (Ed.). (2010). *Multivariate data analysis* (7th ed). Pearson Prentice Hall.

Im, G.-H., Shin, D., & Cheng, L. (2019). Critical review of validation models and practices in language testing: Their limitations and future directions for validation research. *Language Testing in Asia*, *9*(1), 14. https://doi.org/10.1186/s40468-019-0089-4

International Test Commission (ITC). (2019). ITC guidelines for the large-scale assessment of linguistically and culturally diverse populations. *International Journal of Testing*, *19*(4), 301–336. https://doi.org/10.1080/15305058.2019.1631024

Kopalle, P. K., & Lehmann, D. R. (1997). Alpha inflation? The impact of eliminating scale items on Cronbach's alpha. *Organizational Behavior and Human Decision Processes*, *70*(3), 189–197. https://doi.org/10.1006/obhd.1997.2702

Liao, Y. F. (2004). Issues of validity and reliability in second language performance assessment. *Studies in Applied Linguistics and TESOL*, *4*(2), 1–4. https://doi.org/10.7916/SALT.V4I2.1595

Nunnally, J. C. (1967). *Psychometric theory* (1st ed.). McGraw Hill.

Quinn, E., Nation, I.S.P., & Millet, S. (2007). *Asian and Pacific speed readings for ESL learners.* English Language Institute Occasional Publication.

Schmidgall, J., & Xi, X. (2022). Validation of language assessments. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics* (1st ed., pp. 1–13). Wiley. https://doi.org/10.1002/9781405198431.wbeal1242.pub2

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment*, *8*(4), 350–353. https://doi.org/10.1037/1040-3590.8.4.350

Sijtsma, K. (2009). On the Use, the Misuse, and the Very Limited Usefulness of Cronbach's Alpha. *Psychometrika*, *74*(1), 107–120. https://doi.org/10.1007/s11336-008-9101-0

Swain, M. (1993). Second language testing and second language acquisition: Is there a conflict with traditional psychometrics? *Language Testing*, *10*(2), 193–207. https://doi.org/10.1177/026553229301000205

Taber, K. S. (2018). The use of Cronbach's alpha when developing and reporting research instruments in science education. *Research in Science Education*, *48*(6), 1273–1296. https://doi.org/10.1007/s11165-016-9602-2

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education*, *2*, 53–55. https://doi.org/10.5116/ijme.4dfb.8dfd

About the Authors

Steven J. Carter (corresponding author), MA TESOL, is an Assistant Professor in the English Language Teaching and Learning faculty at Brigham Young University–Hawaii. He has taught English language courses in both intensive English and community English programs. He has worked on three major curriculum development projects and has been involved in the creation, analysis, and revision of numerous assessments. His research interests include second language reading and assessment. Email: steven.carter@byuh.edu; Orcid: https://orcid.org/0000-0003-1031-9624.

Matthew P. Wilcox, Ph.D., is the Associate Director for Measurement and Evaluation at the Center for Language Studies at Brigham Young University. He is responsible for the development and maintenance of proficiency-based and achievement language tests at BYU for both major and less-commonly taught languages. His interests include measurement, psychometrics, and data science. Email: wilcoxmp@byu.edu; Orcid: https://orcid.org/0000-0002-6020-529X.