Studies in Second Language Acquisition (2025), **47**, 388–416
doi:10.1017/S0272263124000627

STATE OF THE SCHOLARSHIP

A meta-analysis of the reliability of second language reading comprehension assessment tools

Huijun Zhao[1,2] and Vahid Aryadoust[2]

[1]Sichuan International Studies University, Chongqing, China and [2]National Institute of Education, Nanyang Technological University, Singapore
Corresponding author: Vahid Aryadoust; Email: vahid.aryadoust@nie.edu.sg

(Received 06 January 2024; Revised 16 September 2024; Accepted 26 September 2024)

Abstract

The present study aims to meta-analyze the reliability of second language (L2) reading assessments and identify the potential moderators of reliability in L2 reading comprehension tests. We examined 3,247 individual studies for possible inclusion and assessed 353 studies as eligible for the inclusion criteria. Of these, we extracted 150 Cronbach's alpha estimates from 113 eligible studies (years 1998–2024) that reported Cronbach's alpha coefficients properly and coded 27 potential predictors comprising of the characteristics of the study, the test, and test takers. We subsequently conducted a reliability generalization (RG) meta-analysis to compute the average reliability coefficient of L2 reading comprehension tests and identify potential moderators from 27 coded predictor variables. The RG meta-analysis found an average reliability of 0.79 (95% CI [0.78, 0.81]). The number of test items, test piloting, test takers' educational institution, study design, and testing mode were found to respectively explain 16.76%, 5.92%, 4.91%, 2.58%, and 1.36% of variance in reliability coefficients. The implications of this study and future directions are further discussed.

Keywords: L2 reading comprehension assessment; moderator analysis; reliability; RG meta-analysis

Introduction

Reliability refers to the consistency of test takers' performance across different conditions to ensure fundamentally consistent assessment records for judgments and inferences based on test scores (Chapelle et al., 2008). The various conditions, such as testing environment, test tasks, and test scoring, could lead to random measurement errors, potentially reducing the reliability of test scores (Brown, 2005). Reliability generalization (RG) is a meta-analytic approach to evaluate the variability of reliability coefficients and disentangle the sources of measurement errors across studies (Vacha-Haase, 1998).

https://doi.org/10.1017/S0272263124000627 Published online by Cambridge University Press

RG assesses the overall score reliability for a given measure and evaluates reliability coefficients of test scores obtained from distinct samples across substantive studies with varying design characteristics to identify the optimal predictor for variability in reliability coefficients. Additionally, an RG study can identify the typical measurement conditions that lead to fluctuations in reliability coefficients for test scores, which can potentially explain lower or higher score reliability (Yin & Fan, 2000).

A growing body of RG studies has investigated the measurement error variances in diverse psychometric scales and inventories within the domains of psychology (e.g., Núñez-Núñez et al., 2022), language learning and listening assessment (e.g., Aryadoust, Soo & Zhai, 2023; Shang, Aryadoust & Hou, 2024; Zhai & Aryadoust, 2024), and others (e.g., Hess, McNab & Basoglu, 2014). In addition, a number of predicting variables, such as subject type, number of instrument items, and standard deviation (SD) of the total scores, were identified in empirical studies across different domains, in alignment with the theoretical underpinnings in the specific field and contextual intricacies in empirical studies (e.g., Aryadoust et al., 2023; Hess et al., 2014; Núñez-Núñez et al., 2022). While there are several reliability meta-analysis studies of second language (L2) listening tests (e.g., Shang, 2024) and the overall L2 research (Plonsky & Derrick, 2016), to our knowledge, no previous studies have conducted a comprehensive meta-analysis specifically focusing on the reliability of L2 reading comprehension tests across diverse linguistic and cultural contexts. With regard to L2 assessment, reading comprehension has been regarded as a crucial factor in determining language learners' proficiency, whether in a first, second, or foreign language (Taylor, 2013). For example, reading is viewed to be a necessary skill for academic success (Grabe & Stoller, 2020), as it allows students to comprehend and engage with complex texts across various subjects. Furthermore, strong reading comprehension skills are essential for effective communication and critical thinking (Snow, 2002). Studies have found that fostering reading skills is crucial for developing higher-level cognitive abilities, including problem-solving and analytical thinking, which are essential for academic and professional success (Medranda-Morales, Mieles & Guevara, 2023). To ensure the robustness of assessments in this language skill, it is important to minimize measurement errors and maximize reliability of test scores. Therefore, in this study, we conduct a systematic investigation to identify and examine reliability and its predicting factors in L2 reading comprehension assessment tools. In the following section, we review the factors that predict reliability in L2 reading comprehension assessment tools.

Predictors of reliability in L2 reading comprehension assessment tools

Based on Weir's (2005) model of reading test validation, potential sources of variance in test scores (Brown, 2005), and recommended coding categories for meta-analyses in general (Lipsey & Wilson, 2001; Wilson, 2019), we identified 21 potential predictor variables that might moderate the commonly reported coefficient alphas for L2 reading comprehension tests across studies. These predictors were organized into three categories: study-related variables, test taker–related variables, and test-related variables.

Study-related predictors

Study design

Based on previous research on the effect of research design on study quality (e.g., Hou & Aryadoust, 2021), it is hypothesized that coefficient alpha, serving as an indicator of

https://doi.org/10.1017/S0272263124000627 Published online by Cambridge University Press

study precision, is influenced by different research designs. Research design in the present study refers to the type of research approach, such as experimental or non-experimental, that was specifically proposed as one of the crucial considerations for L2 research meta-analysis (Oswald & Plonsky, 2010).

Study context

Study context (e.g., English as a second language [ESL], English as a foreign language [EFL]) was included as another potential predictor consistent with the assumption that L2 learning is contingent on the social and contextual variables where learning occurs (Gass et al., 2013) and it is also a standard approach in earlier meta-analytical research focusing on reliability estimates (Watanabe & Koyama, 2008).

Test scores

Total scores achieved in L2 reading comprehension tests, along with the mean and SD of test scores, constituted the three additional predictors. Given the variance in total scores across different tests, standardization was necessary, requiring the transformation of raw scores into z-scores. This required including both the mean of the test scores and their total scores. These variables have been empirically investigated in several past RG studies (e.g., Núñez-Núñez et al., 2022).

Sample size

Sample size has been identified as the most commonly used predictor variable in prior RG studies (e.g., Vacha-Haase & Thompson, 2011). Reliability estimates could fluctuate in tandem with sample size; however, if test takers are of a homogenous nature in terms of language proficiency, the variability in the sample would determine whether a larger sample size could exert a positive or negative effect on reliability (Plonsky & Derrick, 2016).

Test taker–related predictors

Variation in reliability coefficients could be due to the characteristics of test participants in studies. Participant heterogeneity, especially in L2 assessment, may arise from individual attributes that are unrelated to language ability (O'Sullivan & Green, 2011). Six test-taker characteristics were identified as predictors potentially contributing to reliability coefficient variability in L2 reading performance, including test takers' age, gender, first language (L1), L2 proficiency level, English learning experience, and educational background.

Age and gender were empirically reported to influence learners' reading comprehension levels. Comprehension abilities vary with age as children develop literacy skills, improving word-level decoding, vocabulary knowledge, and lexical representations (Peng et al., 2018). Gender influences learners' text preferences and engagement, affecting reading behavior and performance (Lepper, Stang & McElvany, 2021). As for test takers' L1, empirical evidence confirms that different language distances between L1 and L2 may cause divergent performance of L2 learners in reading assessments (e.g., Melby-Lervåg & Lervåg, 2014). Other factors, such as cultural influences on comprehension processes (Verhoeven & Perfetti, 2017) and disparities in

https://doi.org/10.1017/S0272263124000627 Published online by Cambridge University Press

educational practices among different language communities (Zhu & Aryadoust, 2020), may also result in variability in reading assessment outcomes. Additionally, test takers' L2 proficiency level was observed to moderate L2 development and test performance in empirical studies. For example, learners with higher proficiency levels were observed to effectively employ reading strategies, resulting in better reading performance than those with lower proficiency levels (McGrath, Berggren & Mezek, 2016). Finally, L2 learners' previous learning experience and educational background could shape their advanced literacy development, as these factors would determine how well L2 learners adjust to the educational practices and literacy expectations in higher-level L2 courses (Grabe & Yamashita, 2022).

Test-related predictors

Test type

Test type, specifically referred to as whether the test items are from standardized tests, from institutional tests, or created by researchers or teachers, may be influenced by macrolevel factors, such as dominant paradigms in language testing and assessment culture, and by microlevel factors, including test writer's cultural background, previous experience, etc. (Shin, 2012). These factors would in turn determine whether the test items capture the specified language abilities of the test takers, thus suggesting whether the obtained test scores precisely reflect the construct being assessed.

Test purpose

Test purpose involves whether the test is for admitting the examinees into a university, placing them into a class, evaluating their progress, or diagnosing their difficulties in learning (Grabe & Yamashita, 2022). High-stakes testing and some achievement assessments involve decisions concerning examinees' future opportunities, thereby being more constrained by concerns of reliability than such low-stakes testing as diagnosis assessment (Grabe & Yamashita, 2022).

Test piloting

Piloting involves trialing test items to ensure item validity and appropriateness of difficulty level for the target test takers (Fulcher, 2013). Test piloting, as a key element in the test design phase, could help improve the validity and reliability of instruments before the actual administration (Grabowski & Oh, 2018). Prior studies have identified a correlation between reliability estimates and piloting status, finding that instruments that were not reported as piloted demonstrated higher reliability than those that were piloted (Plonsky & Derrick, 2016; Sudina, 2021, 2023). Given these findings, piloting status of reading comprehension instruments was postulated as a predictor variable to further explore its impact on reliability.

Test time limit

Time constraints in reading assessments serve as a crucial factor in evaluating the development of automaticity and comprehension (Alderson, 2000), where constrained test time may affect test takers' performance (Weir, 2005). Regarding reading

https://doi.org/10.1017/S0272263124000627 Published online by Cambridge University Press

performance, empirical studies suggested that the time duration allocated for test taking can affect test takers' ability to demonstrate their comprehension skills (Martina, Syafryadin, Rakhmanina & Juwita, 2020). The impact of time constraints has also been investigated in previous meta-analyses regarding reliability, suggesting time constraints have more correlation with interrater reliability than internal consistency (e.g., Plonsky & Derrick, 2016). To further investigate the effect of time constraints on coefficient alpha, we included this variable as a potential predictor.

Test format

Test format, such as multiple choice and other response formats, has been discussed concerning its effects on predicting test takers' reading performance in previous empirical research (e.g., Lim, 2019) and meta-analytical studies (e.g., In'nami & Koizumi, 2009). Given that test format tends to affect test scores unpredictably, it has been conceived as a potential source of variances unrelated to the intended construct (Alderson, Clalpham & Wall, 1995). It is therefore postulated to influence the reliability of reading test scores, since it can affect the difficulty level of the test, which in turn impacts the amount of measurement error and thus the reliability of the test scores.

Testing mode

Given that reading medium could affect the processing of the text (Alderson, 2000), the effects of paper-and-pen–based and computer-based tests on reading comprehension outcomes have been extensively discussed in empirical studies. Empirical evidence suggested that emergent digital reading in L2 is not simply a binary opposition to print reading but rather an extension of it, influenced by the characteristics of digital reading environments, tasks, and readers (Reiber-Kuijpers, Kral & Meijer, 2021). These extraneous factors would potentially affect L2 reading test performance when digital reading is applied in reading assessment. Additionally, other factors in testing mode were also observed to influence reading test outcomes, such as computer familiarity (Chan, Bax & Weir, 2018), mode preference (Khoshsima, Hosseini & Toroujeni, 2017), and digital reading habits (Støle, Mangen & Schwippert, 2020). Considering this, the testing mode was posited to be a potential source of measurement error, thus possibly affecting reliability of reading comprehension instruments.

Cognitive levels

According to Weir's cognitive model of reading (Khalifa & Weir 2009), competent reading requires readers to engage in various cognitive skills; accordingly, reading tests need not solely assess knowledge of information but also evaluate test takers' mastery of the cognitive processes involved in applying that knowledge. Whether reading tasks tap into local-level or global-level comprehension would involve different cognitive processes, thereby affecting reading performance (Liu, 2021). If the cognitive processes targeted by reading tests are either too easy or too challenging for the test-taking sample, this misalignment can significantly reduce the reliability of the test scores by introducing significant sources of measurement error (Fulcher, 2013; Jones, 2012).

https://doi.org/10.1017/S0272263124000627 Published online by Cambridge University Press

Test length

Two variables, the number of test items and the length of reading text in reading comprehension tests, have been posited as salient factors that affect reliability estimates of measurement (Fulcher, 2013). An increase in test item numbers could cause the corresponding increase in reliability, as mathematically predicted by the Spearman-Brown formula and further confirmed by previous RG studies (e.g., Aryadoust et al., 2023). Furthermore, text length, a proxy for reading load, could potentially explain reliability variation in reading comprehension measurement outcomes, as longer text in reading passages would impose more cognitive load on examinees and hence affect their reading performance (Green, Ünaldi & Weir, 2010). Text length investigated in this study involves the number of texts included in the reading comprehension tests.

The present study

The present study aims to conduct an RG meta-analysis to estimate the average reliability estimate from studies involving L2 reading comprehension tests and explore potential moderators explaining variation in the reliability coefficients. The following research questions are posed to fulfill the research purposes:

1. What is the average reliability coefficient of L2 reading comprehension tests?
2. What are the potential moderators of the reliability of L2 reading comprehension tests?

We note that meta-analytic studies integrate effect sizes from diverse research contexts, participants, and instruments, operating under the premise that the construct of interest remains consistent across various measurement tools (Lipsey, 2019). This approach enables a comprehensive assessment of overall effect size and construct reliability. In our study, we aim to apply this principle to reading comprehension tests, assuming that despite the variety of instruments, they all measure the reading comprehension construct, albeit in different contexts and with different test takers. We further note that instrument and test taker characteristics can potentially introduce construct-irrelevant variance and sources of measurement error, as highlighted by Purpura (2004) and Messick (1995). These characteristics are critical in determining test score reliability, with Thompson (1994) noting that participant homogeneity or heterogeneity significantly influences reliability outcomes. Recent RG studies, such as those by Sen (2022) and Núñez-Núñez et al. (2022), have further explored the impact of sample characteristic variables on reliability, underscoring the importance of considering these factors in meta-analytic research on reading comprehension assessments. Thus, our study will synthesize data from multiple research efforts, focusing on how different test taker characteristics influence the reliability of these assessments. This analysis will contribute to the broader discussion on the universal applicability of reading comprehension tests and offer some insights into their reliability across different test taker groups and conditions.

Method

Literature search

The literature search in the present study was conducted assuming that journals represent the major means of dissemination in L2 research rather than book chapters or other publications (Plonsky & Derrick, 2016). Additionally, this study aims to obtain a general overview of reliability reporting practices in L2 reading assessments; therefore, top-tier journals, as a parameter of study quality (Ada, Sharman & Balkundi, 2012), were selected to extract representative samples in this domain.

Accordingly, we restricted the literature search to 55 top-tier, peer-reviewed journals that publish research on applied linguistics and L2 learning, teaching, and assessment, drawing from Zakaria and Aryadoust's (2023) study (see Appendix for journal list). This exclusive focus, admittedly, might generate selection and publication bias or file-drawer problem (Field & Gillett, 2010). To alleviate potential bias and maximize possible coverage in the pertinent domains, we also included six major reading journals (see Appendix) and studies included in meta-analyses related to L2 reading and reliability (In'nami et al., 2022; Shin, 2020; Zhang & Zhang, 2022) in the literature search.

We followed Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) 2020 guidance (Page et al. 2021) and database selection guidelines (In'nami & Koizumi, 2010) to extend the specified journal search to four databases germane to the field: Scopus, the Web of Science, Education Resources Information Center, and Linguistic and Language Behavior Abstracts.

We searched the aforementioned databases with a combination of the key terms "reading test" OR "reading assess*", "reading comprehen*" OR "reading abilit*", "second language", "foreign language", L2, and "bilingual*", in light of Shin's (2020) meta-analysis. No time range was set, but considering comprehensibility, the language was limited exclusively to English. The search procedure was conducted on March 5, 2024, yielding a total of 3,247 articles. After cross-checking titles and abstracts, we removed duplicate results (n = 1,569), publication news (n = 10), and studies not in English (n = 9), leaving 1,659 articles for further scrutiny.

Inclusion and exclusion criteria

We downloaded and reviewed the full text of the 1,659 articles in accordance with the following inclusion and exclusion criteria: (a) the study employed a quantitative methodology; (b) the study involved L2 reading comprehension test at the passage level; (c) the study collected data from L2 learners; (d) the study encompassed English reading comprehension tests; and (e) the study reported information on reliability estimates and sample size of test takers and included more than two items in the test to meet the basic statistical prerequisites.

After removing primary studies that did not meet the previous five inclusion and exclusion criteria, as well as studies with duplicate data (n = 4) and unretrievable studies (n = 8) from conference proceedings, we thoroughly examined the full text of the 353 eligible articles with a focus on reliability estimates. A further 240 articles were excluded for the following reasons: (a) studies (n = 139) applied reliability indices other than Cronbach's alpha; (b) studies (n = 10) reported ranges of coefficient alpha; (c) studies (n = 11) reported coefficient alpha for the test battery rather than exclusively for the reading comprehension test in the battery; (d) studies (n = 68) provided inducted alpha estimates; and (e) studies (n = 12) reported coefficient alphas from the pilot test.

Finally, the rest of the 113 articles with 150 reliability coefficient reports were extracted for the subsequent coding and analysis (see Supplementary Material). The PRISMA flowchart (Moher et al., 2009) in Figure 1 demonstrates the procedure of literature search and data screening.

https://doi.org/10.1017/S0272263124000627 Published online by Cambridge University Press

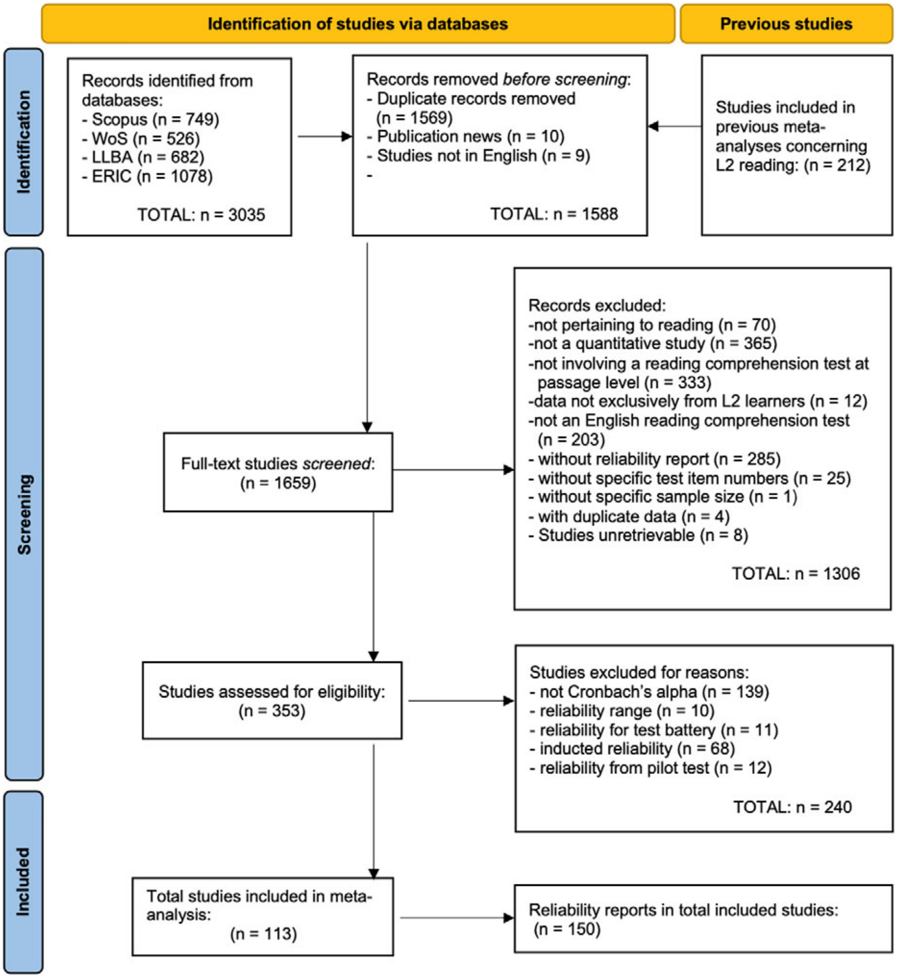
Identification of studies via databases
Previous studies
Identification
Screening
Included
Records identified from databases:
- Scopus (n = 749)
- WoS (n = 526)
- LLBA (n = 682)
- ERIC (n = 1078)
TOTAL: n = 3035
Records removed *before screening*:
- Duplicate records removed (n = 1569)
- Publication news (n = 10)
- Studies not in English (n = 9)
-
TOTAL: n = 1588
Studies included in previous meta-analyses concerning L2 reading: (n = 212)
Full-text studies *screened*: (n = 1659)
Records excluded:
-not pertaining to reading (n = 70)
-not a quantitative study (n = 365)
-not involving a reading comprehension test at passage level (n = 333)
-data not exclusively from L2 learners (n = 12)
-not an English reading comprehension test (n = 203)
- without reliability report (n = 285)
- without specific test item numbers (n = 25)
- without specific sample size (n = 1)
- with duplicate data (n = 4)
- Studies unretrievable (n = 8)
TOTAL: n = 1306
Studies assessed for eligibility: (n = 353)
Studies excluded for reasons:
- not Cronbach's alpha (n = 139)
- reliability range (n = 10)
- reliability for test battery (n = 11)
- inducted reliability (n = 68)
- reliability from pilot test (n = 12)
TOTAL: n = 240
Total studies included in meta-analysis: (n = 113)
Reliability reports in total included studies: (n = 150)

Figure 1. PRISMA flowchart of literature selection process.

The coding scheme

Based on the aforementioned theoretical framework comprising the 21 predictor variables, we designed a coding scheme to capture the characteristics of the primary studies and the potential moderators affecting the reliability coefficients of L2 reading comprehension tests. In the development of the coding scheme, we piloted, revised, and refined codes for variables through an iterative process. The included primary studies were coded in line with the 21 identified potential predictor variables. Along with alpha value and five study descriptors, a total of 27 variables for coding were organized into four categories as study-related variables, test taker–related variables, test-related variables, and observed reliability (see Table 1).

In addition, we also coded for the variable cognitive levels reported by the primary studies. However, 58 out of 150 (38.67%) reliability reports did not document the cognitive levels of the test tasks, which exceeded the minimum of missing values

(10% of the data points) in statistical analysis. Similarly, the variable of time constraint was not reported in 39 out of 150 data points (26%) regarding whether a time limit was imposed on the reading comprehension test. To ensure more accurate statistical analysis with fewer missing values (10%), we ultimately excluded these two potential predictors.

Study characteristics

Overall, 150 Cronbach's reliability coefficient reports from 113 primary studies were coded for subsequent analyses. A majority of the studies were journal articles ($n = 107$), followed by unpublished dissertations ($n = 5$) and book chapters ($n = 1$). Among the journals, *Language Testing* has published the highest number of research ($n = 20$) in L2 reading comprehension assessments, followed by *Reading and Writing* ($n = 9$),

Table 1. Variables, codes, and descriptions of coding scheme

Variables	Values	Descriptions
1. Study related		
Author(s)	Open	Study author(s)
Title	Open	Title of the study
Year	Open	Year of the study publication
Journal source	Open	Title of the journal or book
Study design	1–2	1 = experimental design; 2 = nonexperimental design
Study context	0–3	0 = not reported; 1 = ESL; 2 = EFL; 3 = ESL and EFL
Total score	Open	The total score of the given test; NR = not reported/clear
Mean	Open	The mean of test scores; NR = not reported/clear
SD	Open	The SD of test scores; NR = not reported/clear
Sample size	Open	Number of participants
2. Test taker related		
Age	Open	The mean of test takers' age; NR = not reported/clear
Gender distribution	Open	Percentage of male test takers; NR = not reported/clear
L1	Open	Native language of the test takers as reported in the study; NR = not reported/clear
Years of English learning	Open	Length of the test takers' English learning experience; NR = not reported/clear
L2 proficiency level	0–4	0 = not reported/clear; 1 = beginner; 2 = intermediate; 3 = advanced; 4 = mixed
Educational institution	0–5	0 = not reported; 1 = primary; 2 = secondary; 3 = tertiary; 4 = language institute
3. Test related		
Test type	0–4	0 = not reported; 1 = standardized test; 2 = institutional; 3 = research developed/adapted; 4 = others
Test purpose	0–4	0 = not reported; 1 = proficiency; 2 = placement; 3 = diagnosis; 4 = achievement
Test piloting	0–1	0 = not reported; 1 = yes
Test time limit	0–2	0 = not reported; 1 = timed; 2 = not timed
Test format	0–4	0 = not reported; 1 = multiple choice; 2 = true/false; 3 = open-ended; 4 = mixed
Testing mode	0–2	0 = not reported; 1 = paper and pencil; 2 = computer
Cognitive level	0–3	0 = not reported; 1 = literal; 2 = inferential; 3 = mixed
Test items	Open	The number of items in the reading comprehension test
Text length	0–2	0 = not reported; 1 = with more than one text; 2 = with only one text
4. Observed reliability		
Cronbach's alpha	Open	Cronbach's value as reported
Alpha ID	Open	Individual Cronbach's alpha identification for each primary study based on the number of reliability estimates

Language Learning (n = 6), and *System* (n = 6). There was a total sample size of 70,292 participants for the studies (n = 113) applying Cronbach's estimates. A large proportion of the primary studies (n = 65) involved participants from tertiary education, constituting 57.52% of the total corpus. In terms of research characteristics, only a small proportion of studies (n = 26) adopted an experimental design, accounting for 23% of all the primary studies. Among the primary studies, 76.11% (n = 86) were conducted in an EFL context, with a majority of participants having Asian languages as their L1s, including Chinese (25.66%), Korean (15.93%), and Farsi (10.62%).

Intercoder reliability

Given the complexity of data retrieval and coding tasks, a second researcher (a master's student in applied linguistics) inspected the data extraction procedures and therein confirmed the precision of the data generation. Intercoder reliability was calculated for randomly selected 24 studies (21.24%), with 33 (22%) reliability estimates in the dataset, yielding an agreement rate of 96.78%. The discrepancies in terms of intercoder agreement were discussed and resolved by consensus.

Statistical analyses

The statistical analyses include the following steps: (a) data preparation: evaluating normality of data and identifying potential outliers with standardized residuals and Cook's distances; (b) fitting a multilevel meta-analytic model incorporating three components: sampling error variance, within-study variance, and between-study variance; (c) assessing heterogeneity across studies through Cochran's Q test and I^2 indices; (d) evaluating publication bias with Egger's regression test and trim-and-fill method; and (e) conducting a moderator analysis with the R^2 index to quantify the observed variances that can be explained by the moderator variables.

Data preparation

Following established meta-analytic practices (Lipsey & Wilson, 2001; Cooper, Hedges & Valentine, 2019), we began with data preparations by checking the normality of Cronbach's coefficients. Skewness and kurtosis values between the range of ±2 and ±7 separately suggest a normal distribution of reliability coefficients in the dataset (Kline, 2016). We also evaluated potential outliers in Cronbach's alphas with standardized residuals and Cook's distances to ensure model fitting. Standardized residuals of the reliability estimate exceeding a threshold of $100 \times (1 - 0.05/[2 \times k])^{\text{th}}$ percentile of a standard normal distribution are considered outliers (Viechtbauer & Cheung, 2010). A Bonferroni correction was applied with a significance level of 0.05 (α = 0.05) for Cronbach's alphas in the model. In addition, the reliability coefficient with Cook's distances surpassing the median plus six times the interquartile range of Cook's distances will be considered an overly influential data point (Viechtbauer & Cheung, 2010).

Model fitting

We fitted a three-level meta-analytical model considering the nonindependency of Cronbach's alphas in the selected studies and the fact that the studies included were conducted in different contexts with a variety of participants, which can contribute to the proportion of between-study variability (Borenstein, Hedges, Higgins & Rothstein, 2009).

https://doi.org/10.1017/S0272263124000627 Published online by Cambridge University Press

Thus, the multilevel model in this study partitioned the reliability coefficients variability into three components: sampling error variance, within-study variance, and between-study variance, accordingly, constructing a level one model (participants), level two model (reliability estimates), and level three model (studies) (Assink & Wibbelink, 2016; Harrer Cuijpers, Furukawa & Ebert, 2022). No transformation of reliability coefficients was applied, which is in accordance with the recommendation from Thompson and Vacha-Haase (2000).

Heterogeneity test

We scrutinized heterogeneity through Cochran's Q test and I^2 indices. A statistically significant p value ($p < .05$) in the Q test indicates that factors beyond sampling error contribute to the variability observed across the primary studies. I^2 indices of approximately 25%, 50%, and 75% were considered low, moderate, and large heterogeneity, respectively (Higgins, Thompson & Deeks, 2003).

Publication bias

To examine the potential publication bias, we applied Egger's regression test (Sterne & Egger, 2005), with a significant p value ($p < .05$) indicating potential publication bias in the included studies. The trim-and-fill method (Duval & Tweedie, 2000) was also utilized to help estimate the impact of the potential missing studies on the meta-analysis due to potential publication bias.

Moderator analysis

We conducted a moderator analysis to examine the degree to which the coded predictors explained the variation in the reliability coefficients. However, missing data on the predictors (Pigott, 2019) and the presence of too many moderators in a meta-analysis (Baker et al., 2009) may result in a higher likelihood of a false-positive result. Thus, following Ihlenfeldt and Rios (2023), we included 12 predictors with missing values less than 10% of reliability estimates in the moderator analysis. To assess the statistical significance of the moderator variables and explicate the residual heterogeneity, we used an improved F statistic (Knapp and Hartung, 2003). The R^2 index was employed to quantify the extent to which the observed variances were accounted for by the moderator variables.

All statistical analyses were conducted with the metafor and dmetar (Viechtbauer, 2010) packages in RStudio for macOS.

Results

The average reliability coefficient

Normality checks of Cronbach's alphas (n = 150) indicated a normal distribution (skewness = –0.54, kurtosis = –0.16, M = 0.79, SD = 0.09). In the fitting of the three-level meta-analytical model, no outlier was identified by the examination of the standardized residuals, as none of the coefficient values was larger than ±3.43, the threshold for a standard normal distribution. However, Cook's distances detected one outlier (Choi, Kim & Boo, 2003) in all Cronbach's alphas across studies. A sensitivity analysis was performed to assess the impact of the outlier. The results showed slight differences in the model fit indices of Akaike Information Criterion (AIC) ($\text{AIC}_{\text{with outlier}} = -313.89$;

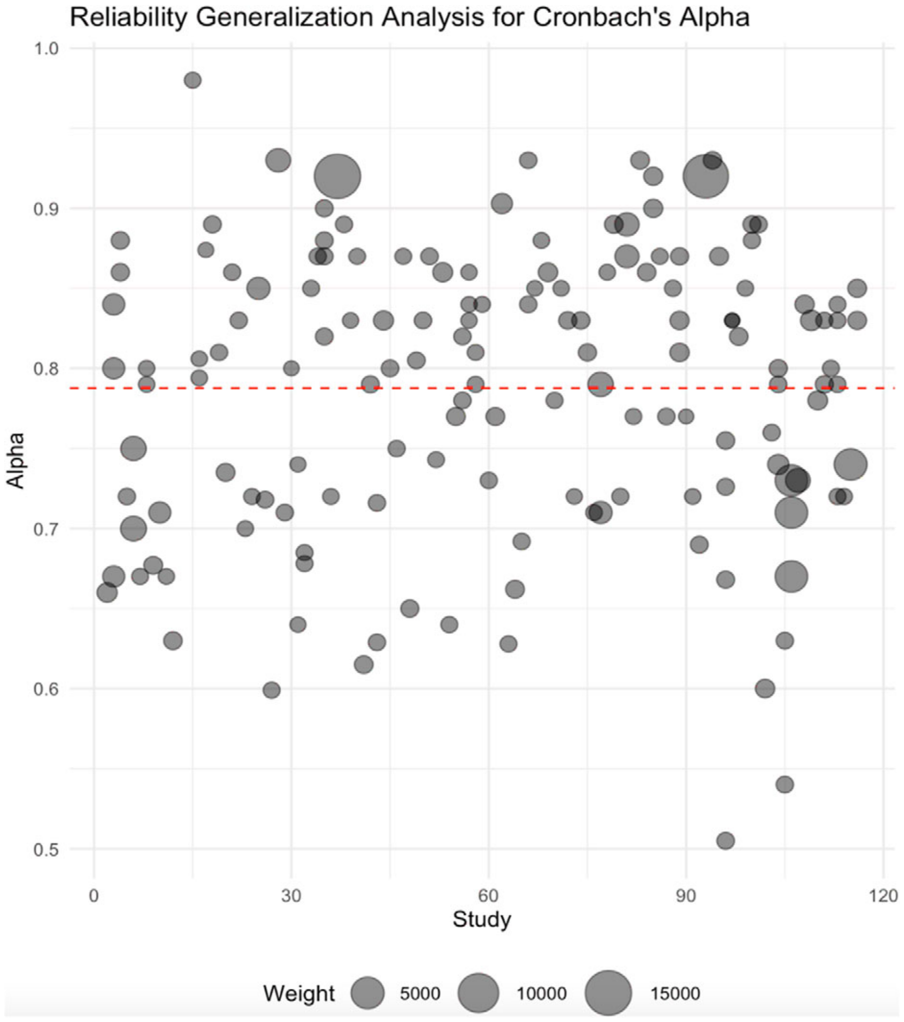

Figure 2. Bubble plot for Cronbach's alphas.

AIC$_{\text{without outlier}}$ = –321.97) and negligible variation in the mean (M) reliability estimates and confidence intervals (M$_{\text{with outlier}}$ = 0.794, standard error [SE]$_{\text{with outlier}}$ = 0.008, 95% CI [0.776, 0.809], $p < .001$; M$_{\text{without outlier}}$ = 0.793, SE$_{\text{without outlier}}$ = 0.008, 95% CI [0.777, 0.810], $p < .001$) in model fit. These results suggest a marginal impact of the identified outliers on the model fitting (Aguinis, Gottfredson & Joo, 2013). Therefore, the raw data were retained for the subsequent analyses without removing the outlier.

A total of 150 Cronbach's coefficients were included in the calculation of the average reliability coefficients, with observed Cronbach's alphas ranging from 0.51 to 0.98. The bubble plot in Figure 2 demonstrates the estimated average Cronbach's value, represented by the red indented line in the upper quadrants of the plot, which is equal to μ = 0.79 (95% CI [0.78, 0.81]). The values of 62 (41.33%) out of 150 reliability coefficients were below the lower bound of the CI.

https://doi.org/10.1017/S0272263124000627 Published online by Cambridge University Press

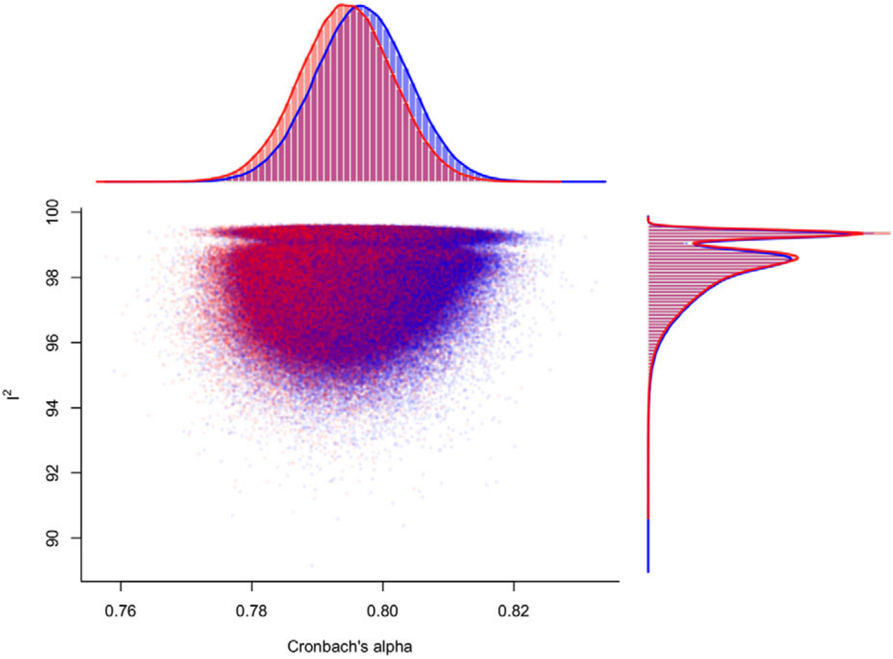

Figure 3. GOSH plot of Cronbach's alphas.

Heterogeneity test

The results of a heterogeneity test (Cochran's Q test, τ^2, and I^2) of Cronbach's alphas suggested significant heterogeneity across studies (Q [149] = 9200.69, $p < .001$, $\tau^2 = 0.01$, I^2 = 98.98%), with 95% prediction interval ranging from 0.65 to 0.95. This result was confirmed by a graphical display of study heterogeneity (GOSH) plot with a normal and symmetric distribution of Cronbach's coefficients at the top and the deviating dotted graph of the I^2 indices below, as displayed in Figure 3. The vertical histogram on the right represents the distribution of the I^2 indices, which is skewed with many of the indices falling roughly above 95%, indicating a large proportion of heterogeneity across reliability coefficients.

To investigate the sources of heterogeneity, the three levels of variance in the meta-analytical model were evaluated individually. Figure 4 presents the total variance distribution across the three levels. The variance in the first level, representing sampling error, constitutes a relatively small proportion of approximately 3.33% of the total variance. A larger amount of heterogeneity variance within studies is observed at level 2, accounting for approximately 23%. The most substantial proportion is found at level 3, where between-study heterogeneity accounts for approximately 73.67% of the overall variation. In the presence of conspicuous between-study variation in the reliability coefficients, it is important to investigate moderators to elucidate potential causes of the variation (Baker et al., 2009).

Publication bias

The trim-and-fill funnel plot in Figure 5 reveals an asymmetric dispersion of Cronbach's coefficients, with a majority of the coefficients located at the upper section of the funnel

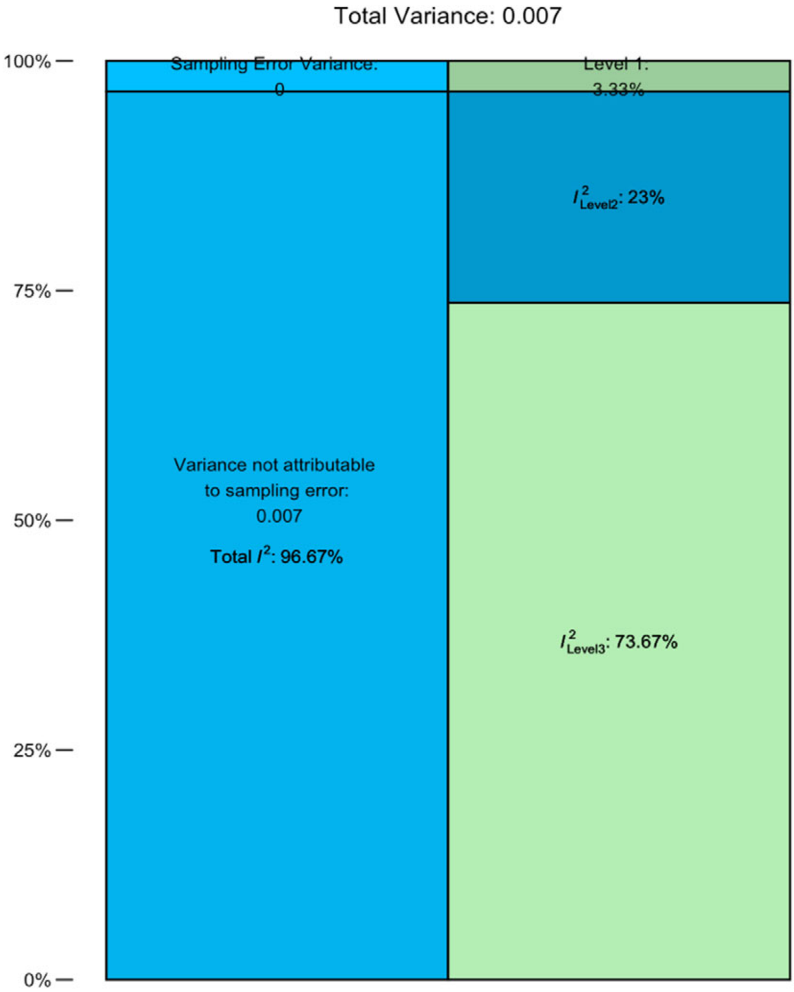

Figure 4. Variance distribution in Cronbach's alphas.

and some sparsely scattered at the left bottom, suggestive of the presence of potential publication bias. The Egger's test substantiated this result with a statistically significant *p* value ($z = -9.32$, $p < .001$). Nonetheless, a trim-and-fill method did not confirm a substantial impact on the meta-analysis, the results of which suggest a possible absence of four studies which, if added, would adjust the mean Cronbach's coefficient from 0.794 to 0.799 and increase variance component from 98. 98% to 99.06%.

Moderator analysis

To assess which predictors of interest might moderate the reliability coefficients of reading comprehension tests, we conducted a moderator analysis for 10 categorical predictor variables and 2 continuous predictor variables. Table 2 presents the results of the omnibus test and post hoc tests, where each category is compared against the

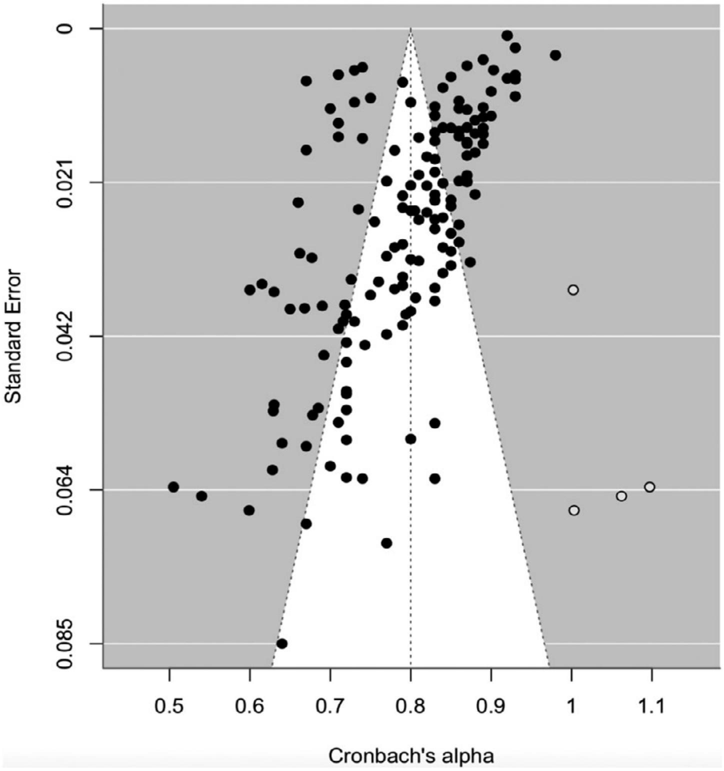

Figure 5. Trim-and-fill funnel plot for Cronbach’s alphas.

anchored category, which is the intercept in this dataset. Overall, 5 out of 12 variables were found to be significant predictors of heterogeneity in the Cronbach’s alpha coefficients, explaining 31.53% of the variance observed. Of the 10 categorical moderators, 4 variables (test piloting, study design, test taker’s educational institution, and testing mode) were found to have a significant moderating effect, together explaining 14.77% of the variance observed. Specifically, test piloting was observed to be a statistically significant moderator of the reliability estimates (F [1, 148] = 11.87, $p < .001$), with an R^2 value of 5.92%, indicating that 5.92% of the between-study variation could be attributed to whether a pilot test was conducted prior to the operationalization. Test taker’s educational institution showed a statistically significant moderating effect (F [3, 145] = 9.65, $p = .02$), too, accounting for 4.91% of the variance between studies, which could be ascribed to the different educational backgrounds of test takers. Study design appeared to be another statistically prominent moderator (F [1, 148] = 5.68, $p = .02$), evidenced by an R^2 value of 2.58%, suggesting that 2.58% of the between-study variance could be explained by whether the primary study adopted an experimental or nonexperimental design. The variable testing mode also emerged as a statistically noticeable moderator (F [1, 148] = 3.98, $p = .05$), accounting for 1.36% of variances among alpha values. This indicates that 1.36% of variances between studies could be attributed to whether the test was computer based or paper based. Other six categorical variables, including study context, text length, test type, test purpose, test format, and test takers’ L1, had weak or no obvious moderating effects.

https://doi.org/10.1017/S0272263124000627 Published online by Cambridge University Press

Table 2. Moderator analysis results for categorical predictor variables.

Moderators	*n*	Estimate	SE	*z* value	*p* value	95% CI	R^2 (%)
Design					.02*		2.58
Experimental	43	0.77	0.01	58.12	<. 001†	[0.74, 0.79]	
Nonexperimental	107	0.04	0.02	2.38	.02*	[0.01, 0.07]	
Context					.17		1.27
EFL	117	0.79	0.01	101.51	< .001†	[0.78, 0.81]	
ESL	29	0.02	0.02	1.03	.31	[–0.02, 0.05]	
EFL and ESL	3	0.08	0.05	1.64	.10	[–0.02, 0.17]	
L1					.91		0
Arabic	6	0.81	0.04	22.21	< .001†	[0.74, 0.88]	
Chinese	37	–0.03	0.04	–0.63	.53	[–0.10, 0.05]	
Dutch	11	0	0.04	0.07	.95	[–0.08, 0.09]	
Farsi	19	0	0.10	0.10	.92	[–0.08, 0.09]	
Japanese	4	–0.03	0.06	–0.60	.56	[–0.15, 0.08]	
Korean	22	–0.01	0.04	–0.24	.81	[–0.09, 0.07]	
Spanish	5	0	0.06	0.06	.95	[–0.10, 0.11]	
Slovenian	4	–0.01	0.06	–0.17	.86	[–0.12, 0.10]	
Turkish	8	–0.05	0.05	–1.12	.26	[–0.14, 0.04]	
Mixed	19	0	0.04	0.02	.98	[–0.08, 0.08]	
Others	15	–0.02	0.04	–0.58	.56	[–0.10, 0.06]	
Educational institution					.02*		4.91
Language institute	12	0.83	0.02	34.75	< .001†	[0.78, 0.88]	
Primary	22	–0.07	0.03	–2.29	.02*	[–0.13, 0.01]	
Secondary	30	–0.01	0.03	–0.21	.83	[–0.06, 0.05]	
Tertiary	85	–0.04	0.03	–1.54	.12	[–0.09, 0.01]	
Test format					.83		0
MC	107	0.79	0.01	95.91	< .001†	[0.78, 0.81]	
T/F	5	0	0.04	0.08	.93	[–0.07, 0.08]	
Open questions	10	–0.01	0.03	–0.30	.77	[–0.06, 0.05]	
Mixed	28	0.02	0.02	0.85	.40	[–0.02, 0.05]	
Test type					.44		0
Institutional	11	0.79	0.03	31.05	< .001†	[0.74, 0.84]	
Standardized	66	0.02	0.03	0.58	.57	[–0.04, 0.07]	
Researcher designed/adapted	67	–0.01	0.03	–0.27	.78	[–0.06, 0.05]	
Others	3	–0.02	0.06	–0.34	.73	[–0.13, 0.10]	
Test purpose					.50		0
Achievement	24	0.78	0.02	44.03	< .001†	[0.75, 0.82]	
Proficiency	115	0.01	0.02	0.61	.54	[–0.03, 0.05]	
Diagnosis	5	0.03	0.04	0.72	.47	[–0.05, 0.11]	
Placement	6	0.06	0.04	1.46	.14	[–0.02, 0.13]	
Testing mode					.05*		1.36
Computer	41	0.77	0.01	56.69	< .001†	[0.75, 0.80]	
Paper and pencil	104	0.03	0.02	1.99	.05	[0, 0.06]	
Test piloting					< .001†		5.92
Not reported	117	0.81	0	107.42	< .001†	[0.79, 0.82]	
Yes	33	–0.06	0.02	–3.45	.001†	[–0.09, –0.03]	
Text length					.17		.71
With more than one text	121	0.80	0.01	104.86	< .001†	[0.78, 0.81]	
With one text	25	–0.03	0.02	–1.38	.17	[–0.07, 0.01]	
Total variance explained							16.75

Note: MC = multiple choice; T/F = true/false.
*p < .05
†p < .001.

We further performed a metaregression for two continuous predictors, sample size and the number of test items, to evaluate their moderating effect on the reliability coefficient outcomes. The sample size variable was observed to have no significant

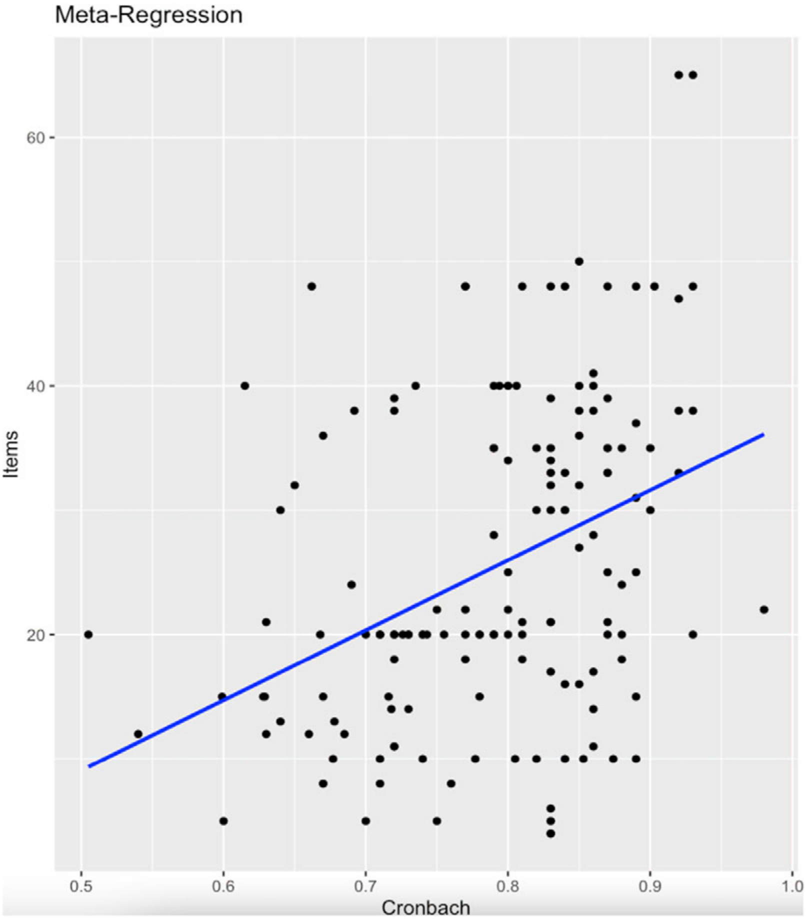

Figure 6. Metaregression of test items.

effect on the coefficients (F [1] = 1.36, $\beta = 0$, SE = 0, $p = .24$), which is confirmed by the R^2 value ($R^2 = 0$), while the number of test items displayed a noticeable moderating effect (F [1] = 26.23, $p < .001$), with a relatively small regression coefficient ($\beta = 0.003$, SE = 0.001) (see Figure 6). This suggests that there is a positive association between the reliability coefficient outcomes and the total number of test items. Overall, 16.76% of between-study variability could be accounted for by the number of test items, as indicated by the R^2 value ($R^2 = 16.76\%$).

Discussion

Average reliability and heterogeneity

The first objective of the present study was to investigate the average reliability coefficients obtained in empirical L2 reading comprehension studies. We obtained an average Cronbach's alpha of 0.79 (95% CI [0.78, 0.81]), which is lower than 0.82 reported in a meta-analysis of all reliability coefficient indices in L2 research based on a larger body of 1,112 Cronbach's alphas (Plonsky & Derrick, 2016). While this average

https://doi.org/10.1017/S0272263124000627 Published online by Cambridge University Press

reliability value is regarded as satisfactory for ability tests (Field, 2018), within the domain of language assessment and educational research, this value would be considered fairly moderate, according to more stringent benchmarks (Brown, 2014; Taber, 2018). It should be noted that low-reliability estimates can result in reduced statistical power and attenuated effect size in substantive research, which would undermine the accuracy and robustness of the research findings (Oswald & Plonsky, 2010; Plonsky, 2013).

It is also noteworthy that 62 out of 150 coefficient alpha values were below the lower bound of the CI, which might attenuate the overall internal consistency of L2 reading comprehension tests included in this RG meta-analysis. Specifically, the lower reliability in these 62 reports may indicate potential issues in the measurement tools or methods they adopted, such as poor item construction or misalignment with the intended construct (Meyer, 2010). The reduced reliability of these studies might stem from factors like the tools' multidimensionality, the clarity and relevance of measurement items, and whether the instrument comprehensively covers the construct (Jones, 2012). To facilitate meta-analytical examinations of the potential causes of low reliability, it is suggested that future authors should include the tests and raw data in their publications.

Relatedly, it was found that the alpha coefficients included in the present study were significantly nonuniform. The result of the heterogeneity test exhibited a large amount of variability among Cronbach's coefficients (I^2 = 98.98%), in which study-level variances accounted for a significant share (73.67%). Given the variation in testing contexts, testing conditions, and participants, it is not surprising that the observed reliability coefficients align with this variability and depend on contextual factors, population characteristics, and other study-specific variables. This, therefore, provides further evidence showing that reliability estimates like Cronbach's alpha are highly susceptible to the specific characteristics and conditions of each study (Zakariya, 2022).

Moderators of reliability coefficients

To answer the second research question, a moderator analysis was conducted on 10 categorical predictor variables and 2 continuous predictor variables. Five variables (the number of test items, test piloting, test takers' educational institution, study design, and testing mode) were found to have a significant moderating effect on coefficient alpha values. Two other variables, study context and text length, appeared to marginally contribute to the explained between-study variability of coefficient alphas, notwithstanding their lack of statistical significance. The remaining five variables—test takers' L1, test type, test purpose, test format, and sample size—had no significant effects on coefficient alphas.

The number of test items emerges as a significant moderator for reliability estimates (F [1] = 26.23, $p < .001$), explaining the largest amount of variance in coefficient alphas (R^2 = 16.76%) across the included primary studies. As expected, the results reveal a positive association between the reliability coefficient outcomes and the total number of test items. This finding is congruent with the prediction of the coefficient alpha formula, suggesting that all other things held constant, an increase in the number of test items will increase reliability coefficient (Taber, 2018; Zhai & Aryadoust, 2024). In the domain of language assessment, as long as the stochastic independence of items (i.e., the independence between-item responses) holds, the increase of test items would increase the reliability of measurements, as each test item contributes to the

https://doi.org/10.1017/S0272263124000627 Published online by Cambridge University Press

information about the test taker's ability (Fulcher, 2013). Previous RG studies for psychometric scales or questionnaires have also found a positive relationship between reliability coefficients and the number of items (e.g., Aryadoust et al., 2023; Sen, 2022). The present study shows that in the context of reading assessment, the alpha coefficient is partially dependent on the number of test items.

Test piloting is another significant moderator of reliability coefficients ($R^2 = 5.92\%$, $p < .001$), accounting for a small amount of heterogeneity between studies. Notably, reading instruments with pilot testing exhibited a negative coefficient (slope) compared to instruments without test piloting reports (the reference group) (see Table 2), indicating that reading instruments that underwent a pilot test tend to have lower reliability estimates compared to those that did not undergo a pilot test. This finding, though surprising, aligns with the findings of previous studies (Plonsky & Derrick, 2016; Sudina, 2021, 2023). Plonsky and Derrick (2016) explained that researchers who did not report piloting their instruments tended to choose reliability coefficients with higher median estimates. Sudina (2023), focusing on Cronbach's alpha, attributed the lower reliability of piloted instruments to limited transparent reporting, where researchers possibly failed to report their pilot testing. To investigate this further, we reexamined our data and found that only 2 out of 25 studies documented alpha values for both the pilot and final tests. Most of the studies reported piloting to confirm the appropriateness of reading materials, test procedures, and test time. These, admittedly, are valid purposes for test piloting (Mackey & Gass, 2021), but most researchers seem to overlook important statistical analyses in piloting, particularly internal consistency. Internal consistency analysis could help researchers ensure the validity of instrument items and identify ways to improve items if necessary (Green, 2020). To avoid the potential pitfalls of low reliability, it is imperative to pilot the instrument in advance of the actual administration (Grabowski & Oh, 2018), but equally important to report the reliability statistics of both pilot and main study test scores.

Test takers' educational institution also demonstrated a significant moderating effect on Cronbach's alphas ($R^2 = 4.91\%$, $p = .02$), where the intercept "language institute" and the category "primary school" were statistically significant. It may be said that the moderating effect of "language institute" stems from the heterogeneity of the test takers from diverse backgrounds and at different proficiency levels. This finding is supported by the literature indicating that reliability can differ when the same instrument is administered to the participants of heterogenous or homogenous nature (Thompson & Vacha-Haase, 2000; Yin & Fan, 2000). Similarly, reliability of test scores for primary school students may be affected by various extraneous factors, including test takers' physiological and psychological variations, test methods, and test design (Papp, 2019). This was evidenced by the findings that the "primary school" category observed a decrease in the coefficient alpha compared to the baseline category (language institute). Further examination of the data in our study revealed that 6 out of 22 reports documented reliability estimates below 0.70.

Study design was also found to moderate coefficient alpha values ($R^2 = 2.58$ %, $p = .02$). The reliability estimates of studies with nonexperimental design, on average, appeared to be higher than those of the experimental design group, with statistical significance (see Table 2). Further examination of the data revealed that most of the included studies (72%) adopted a quasi-experimental approach without a pretest, control group, or random assignment. A quasi-experimental approach, though useful, might introduce confounding noise, potentially lowering reliability estimates (Gliner, 2017). It is noteworthy that an even larger proportion of the included studies utilized nonexperimental designs, predominantly with intact groups, which would potentially

https://doi.org/10.1017/S0272263124000627 Published online by Cambridge University Press

compromise methodological rigor of the studies (Plonsky & Gass, 2011). This practice was identified as normative in L2 research possibly due to the prevalence of classroom-based research, where logistical challenges or ethical problems may exist in abundance (Plonsky, 2013).

Testing mode also appeared to be a statistically significant moderator of Cronbach's coefficients (R^2 = 1.36%, p = .05), where computer-based tests—when compared with the intercept—appeared to be more likely to moderate coefficient alphas than paper-based tests. Previous studies have shown that digitally delivered information can result in subtle changes in test takers' reading behavior, including the time taken to complete tasks, patterns of eye movement, and self-evaluation of performance (Pengelley, Whipp & Rovis-Hermann, 2023). These changes could possibly lead to inconsistency of test scores under on-screen testing conditions. This finding further underscores the importance of test piloting, especially when introducing a new testing mode.

Two other variables, study context and text length, explained a small proportion of between-study variability of coefficient alphas, although their moderating effect on coefficient alphas was not statistically significant. Regarding the study context (R^2 = 1.27%, p = .17), although not generally significant, the EFL condition had a greater likelihood of influencing coefficient alphas compared to ESL condition across the studies. This distinction might be attributed to variations in language performance due to differing learning contexts. L2 learning contexts, whether in an EFL or ESL context, would influence the quality and quantity of learners' input, output, and interactions, thereby generating distinctive L2 developmental patterns and skills (Yu, Janse & Schoonen, 2021). In addition, L2 reading performance may also be contingent on the miscellaneous effect of social influences and individual differences (Prater, 2009). Thus, the reading performance of L2 learners in an EFL context might exhibit greater variability when compared with their counterparts in an ESL context, but the reliability of test outcomes may not be affected by study context, as suggested by the evidence in this study.

Text length was found to explain a minimal and statistically nonsignificant amount of between-study variability (R^2 = .71%, p = .17), where instruments with more than one text seemed to be more inclined to moderate coefficient alphas. Evidence from prior studies suggested that longer texts can affect test takers' cognitive processing (Green et al., 2010) and increase test takers' unintentional disengagement (Forrin et al., 2021), potentially resulting in variances in test scores. However, the results of this study suggest that the number of texts included in the L2 reading tests does not directly impact the reliability of test scores. Indeed, other text-related factors, such as text readability, text imageability, text genre and intertext relationship, merit attention regarding their relationship with the reliability of test outcomes. We intended to explore these factors initially but were hindered by inconsistent evaluation tools and reporting practices or insufficient reporting of text features in the primary studies. Therefore, we advocate for future research to include these text features to enhance replicability of investigations and enable exploration of their interactions.

The remaining five variables—test takers' L1, test format, test type, test purpose, and sample size—had no discernible moderating effect on coefficient alphas and did not account for the variances of coefficient alphas across studies. Test takers' L1 (R^2 = 0, p = .91) was not found to explain the variability in moderate coefficient alpha values. Upon finding this result, we further examined L1 and associated alpha values reported in the primary studies, particularly L1–L2 language and script distance, to identify any possible patterns not captured by the omnibus test. However, no noticeable pattern

https://doi.org/10.1017/S0272263124000627 Published online by Cambridge University Press

emerged, suggesting that test takers' L1 is not likely to contribute to the reliability of test scores.

Test format was hypothesized to potentially affect reliability estimates, considering reading questions presented in different formats may assess distinct reading componential skills (Lim, 2019). However, the results of the study ($R^2 = 0$, $p = .83$) did not demonstrate a positive moderating effect. This might be explained by the absence of identified test format effects in L2 reading research (In'nami & Koizumi, 2009). While varied test formats can lead to different item-response processing, as long as the test is "valid," the extracognitive processes involved in reading test items due to test methods that are unrelated to the construct being measured and thus would not significantly affect the variance in test scores (Lim, 2019). The findings of this study indicate that test format in reading instruments may have less likelihood of affecting the reliability of test outcomes.

Test type ($R^2 = 0$, $p = .44$) and test purpose ($R^2 = 0$, $p = .50$), two interconnected variables, did not appear to account for variation in coefficient alphas. It was hypothesized that standardized tests might moderate reliability coefficients differently compared to classroom-based tests, as the former encapsulate a range of component reading abilities and include a variety of tasks tailored to participants at varying proficiency levels (Grabe, 2009). Similarly, high-stakes tests and some achievement assessments, which influence examinees' future opportunities, were posited to be more constrained by reliability concerns than low-stakes tests like diagnostic assessments (Grabe & Yamashita, 2022). However, the results did not substantiate these postulations. This might be attributed to the likelihood that whether standardized or classroom based, high-stakes tests or low-stakes tests exhibit comparable levels of consistency in assessing reading abilities. This finding suggests that the diversity of reading tasks in standardized tests and high-stakes tests may not necessarily result in higher reliability compared to classroom-based tests and low-stakes tests, as reflected in the data analyzed for this study.

Finally, sample size, albeit a crucial concern in study design, did not appear to affect coefficient alphas ($F\,[1] = 1.36, \beta = 0, p = .24$). This outcome may be explained based on alpha's formula $\alpha = \frac{N \cdot \bar{c}}{\bar{v} + (N-1) \cdot \bar{c}}$, wherein the effect of sample size is notably absent in directly affecting the magnitude of coefficient alpha (Peterson, 1994). This observed lack of substantive correlation between sample size and coefficient alpha resonates with the findings from prior studies (e.g., Aryadoust et al., 2023; Sen, 2022).

Limitations and future studies

Although the present study contributes to theoretical and methodological advances in language assessment, it is subject to several limitations. An arguable limitation is the exclusion of unpublished studies and other studies from the unselected journals in this RG study. This involves a classic trade-off in meta-analysis: prioritizing inclusiveness or study quality. The present study adopts a quality-first approach, given that the reliability coefficient is correlated with the quality of the primary study (for detailed rationales, see Norris & Ortega, 2006; Oswald & Plonsky, 2010). Future research could consider incorporating these unpublished or unselected studies into the RG study and compare the outcomes with those obtained in the present study.

Another potential limitation is the omission of some potential predictor variables due to insufficient reporting in the primary studies. These include test scores, test takers' age, gender, proficiency level, time constraints, and some text-specific characteristics, such as cognitive levels, which had to be excluded as a result of missing data in the primary studies. Specifically, we tried to code for the variables of cognitive levels and time constraints but failed to include them in the final computational analysis due to too many missing values. The exclusion of these predictor variables could lead to a less comprehensive understanding of the relationship between reliability and its potential influencing factors. Future authors are urged to provide demographics of test takers and full texts of the reading instruments to promote transparency and enhance the potential for prospective meta-analysis research. The scope of this study is also limited by the use of Cronbach's coefficients as the sole measure of reliability. While Cronbach's alpha is the most commonly used index for reliability estimate in our data pool, it is worth noting that coefficient alpha has faced criticism regarding its failure to meet the assumptions (e.g., Kline, 2016; Sijtsma & Pfadt, 2021; Teo & Fan, 2013). Alternative reliability coefficients are recommended in lieu of coefficient alpha, such as coefficients omega, coefficient theta, coefficient H, and the greatest lower bound (for details, see McNeish, 2018; Teo & Fan, 2013). We suggest future studies in L2 reading assessments select alternative reliability coefficients based on their research objectives and data characteristics.

Conclusion

The present study determined the average reliability of L2 reading assessments' Cronbach's coefficients and recognized the number of test items, test piloting, test takers' educational institution, study design, and testing mode as potential moderators explaining 31.53% of variance in the reliability coefficients of L2 reading comprehension tests across the studies.

The present study has important implications for researchers and practitioners in L2 reading assessment in terms of theoretical understanding of reliability and validity as well as empirical research design and test development. First, applied researchers are encouraged to assess and report reliability estimates for each application of a given test and tailor their research design to maximize score reliability of L2 reading assessments. Relatedly, as reliability can be affected by the quality of the research, instrument dimensionality, item characteristics, alignment with the intended construct, and coverage of the intended construct, it is advisable for future researchers to provide this information to improve transparency in L2 reading assessment research. Incorporating these variables into moderator analyses in future meta-analytic studies will further lead to a more thorough understanding of reliability in reading assessment. Second, given the limitations of Cronbach's alpha, researchers could consider using alternative reliability coefficients to achieve more precise measurement outcomes in reading assessment. Third, in test development, it is important to consider the number of test items, test piloting, test takers' educational institution, study design, and testing mode when devising L2 reading assessments. A well-balanced number of items ensures that the test covers the reading skills being assessed, without causing fatigue or disengagement or without hampering the precision of the test. Additionally, piloting the test with a representative sample can help identify potential issues in the test items to ensure validity and reliability of the final version of the test. Careful attention to the testing mode, whether digital or print, is also crucial, as it can influence test takers' reading assessment outcomes. Finally, and importantly, increased training, along with rigorous standards, is essential for L2 researchers, teachers, and test developers regarding the understanding and application of reliability. Our study provides further evidence for insufficient reporting and knowledge concerning psychometric features of instruments among L2 researchers and practitioners, as discussed by other researchers (e.g., Plonsky and Derrick, 2016). Therefore, we also advocate for comprehensive and systematic training programs by researcher trainers for L2 researchers, teachers, and test developers with respect to instrumentation knowledge. Additionally, more stringent requirements, particularly concerning reliability estimates, by journal reviewers and editors may also possibly foster enhanced practices among L2 researchers. It is hoped that this study will contribute to the existing knowledge in the field, inspire further research, and provide insights for future applications.

Supplementary material. The supplementary material for this article can be found at http://doi.org/10.1017/S0272263124000627.

Data availability statement. The experiment in this article earned Open Materials badges for transparent practices. The data used in this study and the supplemental materials are available from the following link: https://osf.io/2p6vd/.

Acknowledgments. We would like to express our gratitude to the Singapore Ministry of Education (MOE) and the Chinese Scholarship Council for providing the first author with an MOE Scholarship (August 2022 intake) to study for a master of arts degree in applied linguistics at the National Institute of Education, Nanyang Technological University. We also extend our thanks to the editor and anonymous reviewers for their insightful comments and valuable suggestions. Additionally, we are grateful to the Sichuan International Studies University Academic Research Fund (grant ID: SISU202414) for partially supporting this research. Some parts of the text have been revised using artificial intelligence to improve clarity.

References

Ada, S., Sharman, R., & Balkundi, P. (2012). Impact of meta-analytic decisions on the conclusion drawn on the business value of information technology. *Decision Support Systems*, *54*, 521–533. https://doi.org/10.1016/j.dss.2012.07.001

Aguinis, H., Gottfredson, R. K., & Joo, H. (2013). Best-practice recommendations for defining, identifying, and handling outliers. *Organizational Research Methods*, *16*, 270–301. https://doi.org/10.1177/1094428112470848

Alderson, J. C. (2000). *Assessing reading*. Cambridge University Press. https://doi.org/10.1017/CBO9780511732935

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge University Press.

Aryadoust, V., Soo, Y. X. N., & Zhai, J. (2023). Exploring the state of research on motivation in second language learning: A review and a reliability generalization meta-analysis. *International Review of Applied Linguistics in Language Teaching*, *62*, 1093–1126. https://doi.org/10.1515/iral-2022-0115

Assink, M., & Wibbelink, C. J. M. (2016). Fitting three-level meta-analytic models in R: A step by-step tutorial. *The Quantitative Methods for Psychology*, *12*, 154–174. https://doi.org/10.20982/tqmp.12.3.p154

Baker, W. L., White M., Cappelleri, J. C., Kluger, J., Coleman, C. I., & Health Outcomes, Policy, and Economics (HOPE) Collaborative Group. (2009). Understanding heterogeneity in meta-analysis: The role of meta-regression. *International Journal of Clinical Practice*, *63*, 1426–1434. https://doi.org/10.1111/j.1742-1241.2009.02168.x

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (Eds.). (2009). *Introduction to meta-analysis*. John Wiley & Sons. https://doi.org/10.4324/9780203067659

Brown, J. D. (2005). *Testing in language programs: A comprehensive guide to English Language assessment*. McGraw-Hill.

Brown, J. D. (2014). Classical theory reliability. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 1165–1181). John Wiley & Sons, Inc. https://doi.org/10.1002/9781118411360.wbcla054

Chapelle, C. A., Enright, M. K., & Jamieson, J. (Eds.). (2008). *Building a validity argument for the test of English as a foreign language*. Routledge.

Chan, S., Bax, S., & Weir, C. (2018). Researching the comparability of paper-based and computer-based delivery in a high-stakes writing test. *Assessing Writing*, *36*, 32–48. https://doi.org/10.1016/j.asw.2018.03.008

Choi, I. C., Kim, K. S., & Boo, J. (2003). Comparability of a paper-based language test and a computer-based language test. *Language Testing*, *20*, 295–320. https://doi.org/10.1191/0265532203lt258oa

Cooper, H. M., Hedges, L. V., & Valentine, J. C. (Eds.). (2019). *The handbook of research synthesis and meta-analysis* (3rd ed.). Russell Sage Foundation. https://doi.org/10.7758/9781610448864

Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, *56*, 455–463. https://doi.org/10.1111/j.0006-341X.2000.00455.x

Field, A. (2018). *Discovering statistics using IBM SPSS statistics* (5th ed.). SAGE.

Field, A. P., & Gillett, R. (2010). How to do a meta-analysis. *British Journal of Mathematical and Statistical Psychology*, *63*, 665–694. https://doi.org/10.1348/000711010X502733

Forrin, N. D., Mills, C., D'Mello, S. K., Risko, E. F., Smilek, D., & Seli, P. (2021). TL;DR: Longer sections of text increase rates of unintentional mind-wandering. *The Journal of Experimental Education*, *89*, 278–290. https://doi.org/10.1080/00220973.2020.1751578

Fulcher, G. (2013). *Practical language testing*. Routledge. https://doi.org/10.4324/980203767399

Gass, S. M., Behney, J., & Plonsky, L. (2013). *Second language acquisition: An introductory course* (4th ed.). Routledge. https://doi.org/10.4324/9780203137093

Gliner, J. A., Morgan, G. A., & Leech, N. L. (2017). *Research methods in applied settings: An integrated approach to design and analysis* (3rd ed.). Routledge.

Grabe, W. (2009). *Reading in a second language: Moving from theory to practice*. Cambridge University Press. https://doi.org/10.1017/CBO9781139150484

Grabe, W., & Stoller, F. L. (2020). *Teaching and researching reading* (3rd ed.). Routledge.

Grabe, W., & Yamashita, J. (2022). *Reading in a second language: Moving from theory to practice* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/9781108878944

Grabowski, K. C., & Oh, S. (2018). Reliability analysis of instruments and data coding. In A. Phakiti, P. De Costa, L. Plonsky, & S. Starfield (Eds.), *The Palgrave handbook of applied linguistics research methodology* (pp. 541–565). Palgrave Macmillan. https://doi.org/10.1057/978-1-137-59900-1_24

Green, A., Ünaldi, A., & Weir, C. (2010). Empiricism versus connoisseurship: Establishing the appropriacy of texts in tests of academic reading. *Language Testing*, *27*, 191–211. https://doi.org/10.1177/0265532209349471

Green, R. (2020). Pilot testing: Why and how we trial. In P. Winke, & T. Brunfaut (Eds.), *The Routledge handbook of second language acquisition and language testing* (pp. 115–125). Routledge. https://doi.org/10.4324/9781351034784

Harrer, M., Cuijpers, P., Furukawa, T. A., & Ebert, D. D. (2022). *Doing meta-analysis with R: A hands-on guide* (1st ed.). CRC Press. https://doi.org/10.1201/9781003107347

Hess, T. J., McNab, A. L., & Basoglu, K. A. (2014). Reliability generalization of perceived ease of use, perceived usefulness, and behavioral intentions. *MIS Quarterly*, *38*, 1–28. https://doi.org/10.25300/MISQ/2014/38.1.01

Higgins, J. P. T., Thompson, S. G., & Deeks, J. J. (2003). Measuring inconsistency in meta analyses. *British Medical Journal*, *327*, 557–560. https://doi.org/10.1136/bmj.327.7414.557

Hou, Z., & Aryadoust, V. (2021). A review of the methodological quality of quantitative mobile-assisted language learning research. *System*, *100*, 1–15. https://doi.org/10.1016/j.system.2021.102568

Ihlenfeldt, S. D., & Rios, J. A. (2023). A meta-analysis on the predictive validity of English language proficiency assessments for college admissions. *Language Testing*, *40*, 276–299. https://doi.org/10.1177/02655322221112364

In'nami, Y., & Koizumi, R. (2009). A meta-analysis of test format effects on reading and Listening test performance: Focus on multiple-choice and open-ended formats. *Language Testing*, *26*, 219–244. https://doi.org/10.1177/0265532208101006

In'nami, Y., & Koizumi, R. (2010). Database selection guidelines for meta-analysis in applied linguistics. *TESOL Quarterly*, *44*, 169–184. https://doi.org/10.5054/tq.2010.215253

In'nami, Y., Hijikata, Y., & Koizumi, R. (2022). Working memory capacity and l2 reading: A meta-analysis. *Studies in Second Language Acquisition*, *44*, 381–406. https://doi.org/10.1017/S0272263121000267

Jones, N. (2012). Reliability and dependability. In G. Fulcher & F. Davidson (Eds.), *Languagetesting and assessment: An advanced resource book* (pp. 350–362). Routledge. https://doi.org/10.4324/9780203181287

https://doi.org/10.1017/S0272263124000627 Published online by Cambridge University Press

Khalifa, H., & Weir, C. J. (2009). *Examining reading: Research and practice in assessing second language reading*. Cambridge University Press.

Khoshsima, H., Hosseini, M., & Toroujeni, S. M. H. (2017). Cross-mode comparability of computer-based testing (CBT) versus paper-pencil based testing (PPT): An investigation of testing administration mode among Iranian intermediate EFL learners. *English Language Teaching*, *10*, 23–32. https://doi.org/10.5539/elt.v10n2p23

Kline, R. B. (2016). *Principles and practice of structural equation modeling*. The Guilford Press.

Knapp, G., & Hartung, J. (2003). Improved tests for a random effects meta-regression with a single covariate. *Statistics in Medicine*, *22*, 2693–2710. https://doi.org/10.1002/sim.1482

Lepper, C., Stang, J., & McElvany, N. (2021). Gender differences in text-based interest: Text characteristics as underlying variables. *Reading Research Quarterly*, *57*, 537–554. https://doi.org/10.1002/rrq.420

Lim, H. (2019). Test format effects: A componential approach to second language reading. *Language Testing in Asia*, *9*, 6. https://doi.org/10.1186/s40468-019-0082-y

Lipsey, M. W. (2019). Identifying potentially interesting variables and analysis opportunities. In H. M. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (3rd ed.) (pp. 141–151). Russell Sage Foundation. https://doi.org/10.7758/9781610448864.4

Lipsey, M. W., & Wilson, D. B. (2001). *Practical Meta-analysis*. SAGE Publications.

Liu, H. (2021). Does questioning strategy facilitate second language (L2) reading comprehension? The effects of comprehension measures and insights from reader perception. *Journal of Research in Reading*, *44*, 339–359. https://doi.org/10.1111/1467-9817.12339

Mackey, A., & Gass, S. M. (2021). *Second language research: Methodology and design* (3rd ed.). Routledge. https://doi.org/10.4324/9781003188414

Martina, F., Syafryadin, S., Rakhmanina, L., & Juwita, S. (2020). The effect of time constraint on student reading comprehension test performance in narrative text. *Journal of Languages and Language Teaching*, *8*, 323–329. https://doi.org/10.33394/jollt.v8i3.2625

McGrath, L., Berggren, J., & Mezek, S. (2016). Reading EAP: Investigating high proficiency L2 university students' strategy use through reading blogs. *Journal of English for Academic Purposes*, *22*, 152–164. https://doi.org/10.1016/j.jeap.2016.03.003.

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods*, *23*, 412–433. https://doi.org/10.1037/met0000144

Medranda-Morales, N., Mieles, V. D. P., & Guevara, M. V. (2023). Reading comprehension: An essential process for the development of critical thinking. *Education Sciences*, *13*, 1068. https://doi.org/10.3390/educsci13111068

Melby-Lervåg, M., & Lervåg, A. (2014). Reading comprehension and its underlying components in second-language learners: A meta-analysis of studies comparing first- and second-language learners. *Psychological Bulletin*, *140*, 409–433. https://doi.org/10.1037/a0033890

Messick, S. (1995). Validity of psychological assessment. *American Psychologist*, *50*, 741–749. https://doi.org/10.1037/0003-066X.50.9.741

Meyer, P. (2010). *Reliability*. New York: Oxford University Press.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & PRISMA Group. (2009). Reprint—Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA statement. *Physical Therapy*, *89*, 873–880. https://doi.org/10.1093/ptj/89.9.873

Norris, J. M., & Ortega, L. (2006). The value and practice of research synthesis for language learning and teaching. In J. M. Norris & L. Ortega (Eds.), *Language learning & language teaching*: *Vol. 13* (pp. 1–50). John Benjamins Publishing Company. https://doi.org/10.1075/lllt.13.04nor

Núñez-Núñez, R. M., Rubio-Aparicio, M., Marín-Martínez, F., Sánchez-Meca, J., López-Pina, J. A., & López-López, J. A. (2022). A reliability generalization meta-analysis of the Padua Inventory-Revised (PI-R). *International Journal of Clinical and Health Psychology*, *22*, 100277. https://doi.org/10.1016/j.ijchp.2021.100277

O'Sullivan, B., & Green, A. (2011). Test taker characteristics. In L. Taylor (Ed.), *Examining speaking: Research and practice in assessing second language speaking* (pp. 36–64). Cambridge University Press.

Oswald, F. L., & Plonsky, L. (2010). Meta-analysis in second language research: Choices and challenges. *Annual Review of Applied Linguistics*, *30*, 85–110. https://doi.org/10.1017/S0267190510000115

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., & Moher, D. (2021). Updating guidance for reporting systematic reviews: Development of

https://doi.org/10.1017/S0272263124000627 Published online by Cambridge University Press

the PRISMA 2020 statement. *Journal of Clinical Epidemiology*, *134*, 103–112. https://doi.org/10.1016/j.jclinepi.2021.02.003

Papp, S. (2019). Assessment of young English language learners. In S. Garton, & F. Copland (Eds.), *The Routledge handbook of teaching English to young learners* (pp. 389–409). Routledge.

Peng, P., Barnes, M., Wang, C., Wang, W., Li, S., Swanson, H. L., Dardick, W., & Tao, S. (2018). A meta-analysis on the relation between reading and working memory. *Psychological Bulletin*, *144*, 48–76. https://doi.org/10.1037/bul0000124

Pengelley, J., Whipp, P. R., & Rovis-Hermann, N. (2023). A testing load: Investigating test mode effects on test score, cognitive load and scratch paper use with secondary school students. *Educational Psychology Review*, *35*, 35–67. https://doi.org/10.1007/s10648-023-09781-x

Peterson, R. A. (1994). A meta-analysis of Cronbach's coefficient alpha. *Journal of Consume Research*, *21*, 381–391. https://doi.org/10.1086/209405

Plonsky, L. (2013). Study quality in SLA: An assessment of designs, analyses, and reporting practices in quantitative L2 research. *Studies in Second Language Acquisition*, *35*, 655–687. https://doi.org/10.1017/S0272263113000399

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., & Gass, S. (2011). Quantitative research methods, study quality, and outcomes: The case of interaction research. *Language Learning*, *61*, 325–366. https://doi.org/10.1111/j.1467-9922.2011.00640.x

Pigott, T. D. (2019). Missing data in meta-analysis. In H. M. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (3rd ed.) (pp. 367– 382). Russell Sage Foundation. https://doi.org/10.7758/9781610448864.4

Prater, K. (2009). Reading comprehension and English language learners. In S. E. Israel, & G. G. Duffy (Eds.), *Handbook of research on reading comprehension* (pp. 607–621). Routledge.

Purpura, J. E. (2004). *Assessing grammar*. Cambridge University Press.

Sen, S. (2022). A reliability generalization meta-analysis of runco ideational behavior scale. *Creativity Research Journal*, *34*, 178–194. https://doi.org/10.1080/10400419.2021.1960719

Reiber-Kuijpers, M., Kral, M., & Meijer, P. (2021). Digital reading in a second or foreign language: A systematic literature review. *Computers & Education*, *163*, 104115. https://doi.org/10.1016/j.compedu.2020.104115

Shang, Y., Aryadoust, V., & Hou, Z. (2024). A meta-analysis of the reliability of second language listening tests (1991–2022). *Brain Sciences*, *14*, 746. https://doi.org/10.3390/brainsci14080746

Shang, Y. (2024). *Reliability generalization meta-analysis of L2 listening tests* [Unpublished M.A. dissertation]. Nanyang Technological University.

Shin, D. (2012). Item writing and writers. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 237–248). Routledge. https://doi.org/10.4324/9780203181287

Shin, J. (2020). A meta-analysis of the relationship between working memory and second language reading comprehension: Does task type matter? *Applied Psycholinguistics*, *41*, 873–900. https://doi.org/10.1017/S0142716420000272

Sijtsma, K., & Pfadt, J. M. (2021). Part II: On the use, the misuse, and the very limited usefulness of Cronbach's alpha: Discussing lower bounds and correlated errors. *Psychometrika*, *86*, 843–860. https://doi.org/10.1007/s11336-021-09789-8

Snow, C. E. (2002). *Reading for understanding: Toward an R&D program in reading comprehension*. Rand.

Støle, H., Mangen, A., & Schwippert, K. (2020). Assessing children's reading comprehension on paper and screen: A mode-effect study. *Computers & Education*, *151*, 103861. https://doi.org/10.1016/j.compedu.2020.103861

Sterne, J. A. C., & Egger, M. (2005). Regression methods to detect publication and other bias in meta-analysis. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment and adjustments* (pp. 99–110). John Wiley & Sons Ltd. https://link.springer.com/10.1007/s11336-006-1450-y

Sudina, E. (2021). Study and scale quality in second language survey research, 2009–2019: The case of anxiety and motivation. *Language Learning*, *71*. https://doi.org.libproxy.nie.edu.sg/10.1111/lang.12468

Sudina, E. (2023). Scale quality in second-language anxiety and WTC: A methodological synthesis. *Studies in Second Language Acquisition*, *45*, 1427–1455. https://doi.org/10.1017/S0272263122000560

https://doi.org/10.1017/S0272263124000627 Published online by Cambridge University Press

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

Taylor, L. (2013). *Testing reading through summary: Investigating summary completion tasks for assessing reading comprehension ability*. Cambridge University Press.

Teo, T., & Fan, X. (2013). Coefficient alpha and beyond: Issues and alternatives for educational research. *The Asia-Pacific Education Researcher*, *22*, 209–213. https://doi.org/10.1007/s40299-013-0075-z

Thompson, B. (1994). Guidelines for authors. *Educational and Psychological Measurement*, *54*, 837–847.

Thompson, B., & Vacha-Haase, T. (2000). Psychometrics is datametrics: The test is not reliable. *Educational and Psychological Measurement*, *60*, 174–195. https://doi.org/10.1177/0013164400602002

Vacha-Haase, T. (1998). Reliability generalization: Exploring variance in measurement error affecting score reliability across studies. *Educational and Psychological Measurement*, *58*, 6–20. https://doi.org/10.1177/0013164498058001002

Vacha-Haase, T., & Thompson, B. (2011). Score reliability: A retrospective look back at 12 years of reliability generalization studies. *Measurement and Evaluation in Counseling and Development*, *44*, 159–168. https://doi.org/10.1177/0748175611409845

Verhoeven, L., & Perfetti, C. A. (2017). *Learning to read across languages and writing systems*. Cambridge University Press.

Viechtbauer, W. (2010). Conducting meta-analyses in *R* with the metafor package. *Journal of Statistical Software*, *36*, 1–48. https://doi.org/10.18637/jss.v036.i03

Viechtbauer, W., & Cheung, M. W. L. (2010). Outlier and influence diagnostics for meta- analysis. *Research Synthesis Methods*, *1*, 112–125. https://doi.org/10.1002/jrsm.11

Watanabe, Y., & Koyama, D. (2008). A meta-analysis of second language cloze testing research. *Second Language Studies*, *26*, 103–133. http://hdl.handle.net/10125/40694

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Palgrave Macmillan.

Wilson, D. B. (2019). Systematic coding for research synthesis. In H. M. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (3rd ed.) (pp. 153–172). Russell Sage Foundation. https://doi.org/10.7758/9781610448864.4

Yin, P., & Fan, X. (2000). Assessing the reliability of Beck depression inventory scores: reliability generalization across studies. *Educational Technology Research and Development*, *60*, 201–223. https://doi.org/10.1177/00131640021970466

Yu, X., Janse, E., & Schoonen, R. (2021). The effect of learning context on L2 listening development: Knowledge and processing. *Studies in Second Language Acquisition*, *43*, 329–354. https://doi.org/10.1017/S0272263120000534

Zakaria, A., & Aryadoust, V. (2023). A scientometric analysis of applied linguistics research (1970–2022): Methodology and future directions. *Applied Linguistics Review*. https://doi.org/10.1515/applirev-2022-0210

Zakariya, Y. F. (2022). Cronbach's alpha in mathematics education research: Its appropriateness, overuse, and alternatives in estimating scale reliability. *Frontiers in Psychology*, *13*, 1074430. https://doi.org/10.3389/fpsyg.2022.1074430

Zhai, J., & Aryadoust, V. (2024). A meta-analysis of the reliability of a metacognitive awareness instrument in second language listening. *Metacognition and Learning*, *19*, 879–906. https://doi.org/10.1007/s11409-024-09392-z

Zhang, S., & Zhang, X. (2022). The relationship between vocabulary knowledge and L2 reading/listening comprehension: A meta-analysis. *Language Teaching Research*, *26*, 696–725. https://doi.org/10.1177/1362168820913998

Zhu, X., & Aryadoust, V. (2020). An investigation of mother tongue differential item functioning in a high-stakes computerized academic reading test. *Computer Assisted Language Learning*, *35*, 412–436. https://doi.org/10.1080/09588221.2019.1704788

https://doi.org/10.1017/S0272263124000627 Published online by Cambridge University Press

Appendix

Journal list (* indicates newly added reading journals)

1. *AILA Review*
2. *Applied Linguistics*
3. *Applied Linguistics Review*
4. *Asian ESP Journal*
5. *Assessing Writing*
6. *Bilingual Research Journal*
7. *Bilingualism: Language and Cognition*
8. *CALICO Journal*
9. *Canadian Modern Language Review*
10. *Chinese Journal of Applied Linguistics*
11. *Computer Assisted Language Learning*
12. *ELT Journal*
13. *English for Specific Purposes*
14. *English Teaching*
15. *Foreign Language Annals*
16. *Innovation in Language Learning and Teaching*
17. *International Journal of Applied Linguistics*
18. *International Journal of Bilingual Education and Bilingualism*
19. *International Journal of Bilingualism*
20. *International Journal of Multilingualism*
21. *International Multilingual Research Journal*
22. *IRAL - International Review of Applied Linguistics in Language Teaching*
23. *Iranian Journal of Language Teaching Research*
24. *ITL - International Journal of Applied Linguistics (Belgium)#*
25. *JALT CALL Journal*
26. *Journal of Asia TEFL*
27. *Journal of English for Academic Purposes*
28. *Journal of Multilingual and Multicultural Development*
29. *Journal of Reading Behavior**
30. *Journal of Research in Reading**
31. *Journal of Second Language Writing*
32. *Language Acquisition*
33. *Language Assessment Quarterly*
34. *Language Awareness*
35. *Language Learning*
36. *Language Learning and Technology*
37. *Language Learning Journal*
38. *Language Teaching*
39. *Language Teaching Research*
40. *Language Testing*
41. *Language Testing in Asia*
42. *Linguistic Approaches to Bilingualism*
43. *Modern Language Journal*
44. *Reading and Writing*
45. *Reading and Writing Quarterly*
46. *Reading Psychology**
47. *Reading Research Quarterly**
48. *ReCALL*
49. *RELC Journal*
50. *Research in the Teaching of English*
51. *Scientific Studies of Reading**

https://doi.org/10.1017/S0272263124000627 Published online by Cambridge University Press

52. *Second Language Research*
53. *Studies in Second Language Acquisition*
54. *Studies in Second Language Learning and Teaching*
55. *Study Abroad Research in Second Language Acquisition and International Education*
56. *System*
57. *Teaching English with Technology*
58. *TESOL International Journal*
59. *TESOL Journal*
60. *TESOL Quarterly*
61. *The Reading Teacher**

Cite this article: Zhao, H., & Aryadoust, V. (2025). A meta-analysis of the reliability of second language reading comprehension assessment tools. *Studies in Second Language Acquisition*, 47: 388–416. https://doi.org/10.1017/S0272263124000627