Meta-Analysis of Coefficient Alpha: Empirical Demonstration Using English Language Teaching Reflection Inventory

Bahrul Hayat[1,*]

[1]Faculty of Psychology, UIN Syarif Hidayatullah Jakarta, Indonesia

*Correspondence: Faculty of Psychology, UIN Syarif Hidayatullah Jakarta, Kertamukti St. No. 5, South Tangerang City 12220, Indonesia. E-mail: bahrulhayat@uinjkt.ac.id

Received: November 1, 2023 Accepted: May 3, 2024 Online Published: May 15, 2024

doi:10.5430/jct.v13n2p231 URL: https://doi.org/10.5430/jct.v13n2p231

Abstract

Cronbach's alpha is a reliability coefficient commonly reported in second language (L2) and English language teaching (ELT) studies. The alpha coefficient provides information on the internal consistency of a measuring instrument. The reported alpha coefficients are obtained from, and apply only to, the research sample. However, the estimation of the alpha coefficient for the population has not received the attention of L2 and ELT researchers. This study aims to provide an overview of the alpha coefficient estimation procedure of a measuring instrument for a population with the reliability generalization method, commonly known as alpha coefficient meta-analysis. An example alpha coefficient meta-analysis study—using empirical data of the 29-item English Language Teaching Reflection Inventory (ELTRI) from 27 independent study samples—was conducted to provide an overview of the procedure for applying the method and the information that needs to be reported from the results of the analysis. The results of the study using a random-effect model show that the population alpha of ELTRI was 0.872, indicating excellent reliability; this is followed by application of a mixed-effect model that shows that article type and means of teaching experience significantly impacted ELTRI reliability. Implications for future research are discussed.

Keywords: Cronbach's alpha, meta-analysis, reliability, reliability generalization, research synthesis

1. Introduction

Reliability—an important concept psychometrics—describes the consistencies of a measuring instrument. The Wilkinson & American Psychological Association (APA) Task Force on Statistical Inference (1999) states, "It is important to remember that a test is not reliable or unreliable. Reliability is a property of the scores on a test for a particular population of examinees. Thus, authors should provide reliability coefficients of the scores for the data being analyzed, even when the focus of their research is not psychometrics." In the field of applied linguistics—especially in the context of language assessment—reliability has become a critical issue, as stated in The Encyclopedia of Applied Linguistics (Chapelle, 2012), journal article reporting standards (Derrick, 2016), and special reports on the reliability of large-scale tests such as the TOEFL iBT (see Educational Testing Service, 2020). However, reliability is an attribute of the score generated from the test and not that of the test itself (Lane et al., 2002). In addition to the validity of the measuring instrument, every researcher should report information about the reliability of the measuring instrument used in his research to highlight the consistency of the scores obtained from the measuring instrument (Wilkinson & APA Task Force on Statistical Inference, 1999).

Mathematically, reliability is defined as the ratio of the true score variance to the observed score variance (Raykov & Marcoulides, 2010). In other words, the reliability coefficient provides information on how much of the variation in a person's score is influenced by the person's actual ability. The higher the reliability, the greater the credibility of the score obtained from the test as representing the individual's true ability. Therefore, reliability—also called consistency and trustworthiness—is the extent to which the score obtained from a measurement process can be trusted (Ahrens et al., 2020).

One of the most popular reliability coefficients is Cronbach's alpha (hereinafter referred to as alpha), an indicator of the internal consistency of a test score (Cho & Kim, 2015). Generally, an alpha coefficient value above 0.70 indicates that a test has good internal consistency or, in other words, that the test is reliable (Nunnally, 1978). Alpha—a single number—makes it easy for researchers to make decisions about the reliability of an instrument (Cortina, 1993). Although alpha is commonly used by researchers in the field of applied linguistics, its interpretation is brief and inadequate (Loewen & Plonsky, 2015; Plonsky & Derrick, 2016).

To further increase the statistical power and robustness of the estimated alpha coefficient as information on the reliability of a research instrument, psychometric experts have developed a method in the form of meta-analysis of the alpha coefficient or other reliability coefficients obtained empirically from various reliability tests (e.g., Bonnett, 2010; Rodriguez & Maeda, 2006; Vacha-Haase, 1998). Meta-analysis is a term used by Glass (1976), which refers to a statistical analysis method to quantitatively synthesize various similar research results so that a more comprehensive conclusion can be drawn. Statistical estimation is carried out from research samples with similar research questions, and the results of statistical analysis are also relatively similar. In other words, meta-analysis is the quantitative statistical analysis of the results of existing analyses (Glass, 1976). The main advantage of the meta-analysis method is the aggregation of information from various similar studies, as opposed to that from a single study, thereby providing more complete information.

When the alpha coefficients from various studies are synthesized with the meta-analysis method, it is known as an alpha coefficient meta-analysis study (for example, Bonett, 2010; Rodriguez & Maeda, 2006) or reliability generalization (RG; Vacha-Haase, 1998). RG is a meta-analytic method developed "to characterize the mean measurement error variance across studies (using a particular instrument) and also the sources of variability of these variances across studies (Vacha-Haase, 1998). Although various studies have stated that RG was developed in 1998, Peterson (1994) conducted a meta-analysis of alpha coefficients earlier using a different technique from RG.

RG is developed on the basis that the test reliability obtained from a study is not an attribute of the population. A similar instrument, when administered to different samples at different times, produces different estimates of reliability. This happens because each study sample influences the high or low measurement error. RG was carried out with the aim of combining the reliability estimation results from various studies using the same measuring instrument to obtain the actual alpha estimation of the population and can be used to explore factors influencing the variability of alpha (Vacha-Haase, 1998; Vassar & Bradley, 2011). Each study was treated as a sample from the same study population.

Henson et al. (2001) state that information based on previous research is not sufficient to prove that an instrument is reliable. Given that there are numerous studies using instruments measuring the same construct that have been adapted into various languages or cultures—and generally, each of these studies reports an alpha coefficient—the RG method can be used to estimate the alpha coefficient of the population so that it can be used as an initial reference for using the instrument.

Based on the literature review, a meta-analysis of alpha coefficients has been carried out on various measurement instruments in the field of social sciences, including applied linguistics. For example, in the context of second language (L2) and English language teaching (ELT), a meta-analysis of the alpha coefficient was pioneered by Plonsky and Derrick (2016) in an article entitled "A Meta-Analysis of Reliability Coefficients in Second Language Research." The study conducted a meta-analysis of the reliability coefficient of 537 articles in the context of second language (L2) measurements (Plonsky & Derrick, 2016). In addition, there are also meta-analyses of IELTS scores (Gagen, 2019), second language cloze testing research (Watanabe & Koyama, 2008), and TOEFL iBT (Biber et al., 2011), as well as a step-by-step guide for researchers (Oswald & Plonsky, 2010).

Additionally, in the context of psychology, a meta-analysis of the alpha coefficient was performed for popular instruments such as the Life Orientation Test (Vassar & Bradley, 2010), Reynolds Adolescent Depression Scale (Vassar & Bradley, 2011), Child and Adolescent Perfectionism Scale (Vicent et al., 2019), Alabama Parenting Questionnaire-9 (Liang et al., 2021), and Creativity Achievement Questionnaire (Yörük & Sen, 2022). Concurrently, in education, a meta-analysis of the alpha coefficient was performed using the Motivated Strategies for Learning Questionnaire (Holland et al., 2018), the Runco Ideational Behavior Scale (Sen, 2021), and the Brief Self-Control Scale (Haktanir et al., 2024). However, a meta-analysis of coefficient alpha of the popular measures in L2 and ELT was still limited, although this method was applied in the most current research across disciplines.

The absence of simple explanations regarding the basic concepts, application procedures, statistical estimates, and what information should be reported in conducting a meta-analysis study is a possible reason for the limited use of meta-analyses of alpha in applied linguistics. Therefore, to bridge this gap, this study aims to briefly explain the concept of alpha, discuss its limitations, and elucidate the concept and procedure for conducting a meta-analysis of coefficient alpha. Furthermore, empirical illustrations are presented to provide an overview of the alpha coefficient meta-analysis procedure and interpretation of the results. This study provides technical information for applied linguistics researchers when conducting a meta-analysis of alpha coefficients.

2. Literature Review

2.1 Coefficient Alpha: Basic Concept and Limitations

The alpha coefficient (Cronbach, 1951) is a reliability coefficient that describes the internal consistency of the test score. In addition, the alpha coefficient is a measure of how homogeneously the items in a test measure the construct we are measuring. In other words, the alpha coefficient provides information on how well a measuring instrument can produce a consistent score when measuring a construct. The formulas used to compute alpha coefficients can be found in various studies (Cronbach, 1951). The alpha coefficient ranges from 0 to 1, where a higher value indicates better internal consistency. The basic formula for the alpha coefficient is:

$$r_u = \left(\frac{n}{n-1}\right)\left(1 - \frac{\sum \sigma_i^2}{\sigma_t^2}\right)$$

where n is the number of items in an instrument, σ_t^2 is the variance of the test score, σ_i^2 is the variance of item i, and $\sum \sigma_i^2$ is the total variance of all items. The criterion commonly used in interpreting alpha coefficients is that a coefficient value greater than 0.70 indicates that a test has good internal consistency. A coefficient value of 0.80 is recommended for research that focuses on theory development, and a value of 0.90 is recommended for making important decisions (for example, TOEFL) (Greco et al., 2018; Nunnally, 1978).

Although Cronbach's alpha is popularly used, social science researchers sometimes overlook several details of the use of alpha coefficients. Several studies have criticized the use of alpha coefficients in social science research (e.g., Cho & Kim, 2015; Hoekstra et al., 2019; Sijtsma, 2009). Cronbach's alpha, which was developed using the classical test theory approach, has assumptions—such as those related to parallel tests and uncorrelated errors—that are difficult to fulfill under real conditions. Cronbach's alpha does not produce accurate estimates when the parallel tests and uncorrelated error assumptions are not met (Raykov & Marcoulides, 2010). Furthermore, Streiner and Norman (1989) state that Cronbach's alpha has several problems. First, the alpha coefficient depends not only on the magnitude of the correlation among items but also on the number of items in the test. Second, the alpha coefficient will be higher when combining the two subtests, even though the subtests measure different dimensions. Third, the alpha coefficient will be higher by including as many items as possible measuring the same indicator (redundancy) with a slightly different item formulation.

For researchers who want to understand the concept of test reliability further, several studies have described in detail the limitations of using Cronbach's alpha and reliability in general (see McNeish, 2018; Raykov & Marcoulides, 2019) so that the use of these coefficients is in line with the appropriate conditions. In this study, the Cronbach's alpha used is the original version which is applied to continuous data, while there is another version, namely Ordinal Cronbach's alpha (Chalmers, 2018; Zumbo et al., 2007) which is not discussed in this study even though the Ordinal Alpha has been used by many previous studies (e.g., Rahayu et al., 2022; Soetjipto et al., 2023; Suryadi et al., 2020).

2.2 Meta-Analysis of Coefficient Alpha: Procedure

The alpha coefficient meta-analysis procedure begins by determining the operational definition of the study population. The criteria for studies to be included (inclusion criteria) and not included in the meta-analysis (exclusion criteria) should be clearly defined. The next step is to examine literature containing research reports for sampling. The search should be as comprehensive as possible and include various sources. Inclusion-exclusion criteria are used to assess whether a study could be included in the meta-analysis. After a number of studies that meet the criteria are collected, the researcher extracts and codes the data. Data extraction is the collection of key information from a study that will be used as data in a meta-analysis. In the context of a meta-analysis of alpha, data extraction includes the alpha coefficient, the sample used, and data related to variables that are considered to influence or relate to the alpha reliability test. After the research sample data are complete, meta-analysis can be performed by calculating the effect size (ES), namely the estimation of the alpha coefficient for the population, homogeneity test, and analysis of the factors that affect the variability of the alpha coefficient.

Meta-analysis of alpha coefficients can be carried out using two approaches: the classical approach (see, Bonnett, 2010; Botella et al., 2010; Rodriguez & Maeda, 2006) and the Bayesian approach (see Brannick & Zhang, 2013; Okada, 2015). This study used the classical approach. Meta-analysis of reliability coefficients with the classical approach generally aims to (a) synthesize alpha coefficients from the study sample to obtain population alpha coefficients, (b) perform homogeneity tests, and (c) analyze moderator variables affecting them. To achieve this, researchers must use appropriate statistical methods to synthesize the reliability coefficient data of the measuring instruments to be tested (Botella et al., 2010).

Mathematically, the objective of the alpha coefficient meta-analysis is to generate an estimate of the alpha coefficient ($\hat{\rho}$) for the population. As discussed earlier, to generate an estimate of the alpha coefficient of the population, the researcher must collect research results using the same measuring instrument and treat the alpha values reported from the studies as an alpha sample. Given that it is almost impossible that the alpha values from the collected studies will be normally distributed, but rather will be negatively skewed (see Peterson, 1994; Rodriguez & Maeda, 2006), the alpha sample needs to be transformed into a form that closely follows a normal distribution and needs to be weighted by a certain procedure. Weighting is generally performed using the inversion of the transformed variance alpha (Okada, 2015; Rodriguez & Maeda, 2006). The transformation aims to fit the alpha coefficients to a distribution that is close to normal. Alpha transformation (T_i) is carried out using the following formula (Miller et al., 2018):

$$T_i = (1 - r_{ai})^{1/3}$$

where r_{ai} is the alpha coefficient of the i-th study. The alpha coefficient estimate for the population ($\bar{T}$) is calculated using the following formula (Rodriguez & Maeda, 2006):

$$\bar{\mathrm{T}} = \frac{\Sigma(\mathrm{w_i T_i})}{\Sigma \mathrm{w_i}}$$

However, to calculate the alpha coefficient for the population ($\bar{T}$), it is necessary to weight each alpha coefficient by the number of samples in its study. The weighted value (w_i) for each study is the variance of $\mathrm{T_i}$ and is calculated using the following formula:

$$w_i = 1/v_i$$

where the variance of the transformed alpha (v_i) is obtained using the following formula:

$$v_i = \frac{18\mathrm{J_i}(\mathrm{n_i} - 1)(1 - \mathrm{r_{ai}})^{2/3}}{(\mathrm{J_i} - 1)(9\mathrm{n_i} - 11)^2}$$

where $\mathrm{n_i}$ is the number of samples in the i-th study and Ji is the number of items in the i-th study (Botella et al., 2010). To be able to produce the confidence interval of $\bar{T}$, it is necessary to compute the variance of $\bar{T}$ with the formula:

$$\mathrm{v.} = 1/\Sigma \mathrm{w_i}$$

When the value of v. has been obtained, the square root of v., which is a standard error of the mean ($\sqrt{\mathrm{v.}}$), will be used to calculate the confidence interval of the population alpha through the formula $\bar{T} \pm z_{\alpha/2}\sqrt{\mathrm{v.}}$ The confidence interval is then transformed back into the original metric of alpha with the formula (Botella et al., 2010; Rodriguez & Maeda, 2006):

$$\hat{\rho} = |1 - \bar{\mathrm{T}}^3|$$

The lower and upper limits of the confidence interval are also transformed using the same formula. After calculating the back transformation, population alpha, and confidence interval, further analysis is needed to test the homogeneity of the sample alpha coefficients against the null hypothesis, which states that there is no significant difference between the alpha coefficients across studies. The Q-statistic (Hedges & Olkin, 1985) is used to test homogeneity with the following formula:

$$Q = \Sigma \frac{(T_i - \bar{\mathrm{T}})^2}{v_i}$$

Mathematically, the Q statistic is the ratio of the variance between studies to the variance within individual studies. This statistic has a chi-square distribution with degrees of freedom of k−1, where k is the number of studies (Cheung, 2015). After the Q-statistic is generated, the next step is to calculate the I^2 statistic (Higgins & Thompson, 2002) using the following formula:

$$I^2 = 1 - \frac{k - 1}{Q}$$

where k is the number of studies in the meta-analysis, and Q is the Q-statistic generated previously. It should be understood that calculating all formulas does not have to be done manually (for example, Microsoft Excel), considering that there are many software tools that can be used to perform a meta-analysis, including a meta-analysis of alpha coefficients.

3. Method

3.1 Empirical Illustration Using ELTRI: Coding Strategy

To explain the procedure for implementing the alpha coefficient meta-analysis, the researcher presents an empirical illustration using the English Teaching Reflection Inventory (ELTRI; Akbari et al., 2010), a foundational instrument for measuring English teachers' reflective teaching. Reflective teaching practice is a recurring theme in teachers' professional development and education (Kurosh et al., 2020). This instrument contains 29 items that measure teacher reflection in a particular ELT context. As reported by its developer, ELTRI has a reliability of 0.91 (Akbari et al., 2010) and is cited by many researchers using ELTRI. Since its publication, ELTRI has been cited 202 times in the Google Scholar database, as well as 57 times and 22 times in the Scopus and Crossref databases, respectively. ELTRI has been adapted and used in various countries (e.g., Xu et al., 2015); however, to the best of our knowledge, no meta-analysis has been conducted of ELTRI. Therefore, in this study, a meta-analysis of the alpha coefficient of ELTRI was deemed appropriate to describe the procedure for implementing the alpha coefficient meta-analysis.

For this meta-analysis, the researchers specified the study criteria to be sampled as follows: (1) The studies must be published in English-language scientific journals; ELTRI studies published in the form of proceedings or dissertations were not included in the analysis; and (2) the alpha coefficient is reported of the entire scale of ELTRI. Using these criteria, the sample of publications in this study amounted to 27 studies published between 2010 and 2021. Of the 27 studies, the researchers collected data on the following: (1) alpha coefficient value, (2) sample size, (3) number of items used, (4) gender ratio, by percentage of female respondents, (5) article type, where research that focuses on testing ELTRI psychometric characteristics (measurement) is coded 1, and research that applies ELTRI to test its relationship with other variables (substantive) is coded 0, (6) mean of ELTRI scores, where a higher mean indicates higher teacher reflectivity; and (7) mean of teaching experience reported in each sample, where a higher mean indicates that the sample consists of a high number of experienced teachers.

3.2 Data Analysis Strategy

In the initial analysis, the model used was a random-effects model, which does not involve moderator variables and only produces transformed alpha and population alpha. However, if the data are heterogeneous, then the mixed-effect model should be used at a later stage by including moderator variables in the model to explain the heterogeneity of the data (Hedges & Vevea, 1998). In this study, the '*metafor*' package in R was used (Viechtbauer, 2010) to perform the two stages of analysis mentioned above.

4. Results

4.1 Population Alpha

The results of the alpha meta-analysis of ELTRI showed that the alpha population ($\hat{\rho}$) was 0.872 with a 95% confidence interval of 0.854–0.890. Referring to the 0.70 criterion, which is the cutoff of the alpha coefficient, it can be concluded that the expected ELTRI reliability value is quite high. In other words, from the 27 research samples, it was found that ELTRI had excellent internal consistency. If the researcher intends to adapt ELTRI to other languages and cultures, it is very likely that the reported internal consistency (alpha) of ELTRI will not be far from $\hat{\rho}$ = 0.872 (see Figure 1).

As shown in Figure 1, the range of confidence intervals for the 27 studies varied widely from very wide to very narrow. The sample size is likely one reason for this. When the sample size of a particular study is small, the confidence interval becomes wider. After the alpha estimation of the population was obtained, the next step was to test the homogeneity of the alpha coefficient of the 27 study samples using the Q-statistic and I^2 statistic. The analysis resulted in a Q-value (df = 26) of 328.682, which was statistically significant ($p < 0.001$). When Q is significant, the null hypothesis is rejected, suggesting that the alpha coefficient data from 27 studies are heterogeneous, as shown in Figure 1. The analysis also yielded a statistical value of I^2 of 0.946, which indicates that 94.6% of the variation was caused by variations between studies and 5.4% by variations within each study. Based on these findings, further analysis is needed to explore the factors that influence the variation between studies using a mixed-effect model.

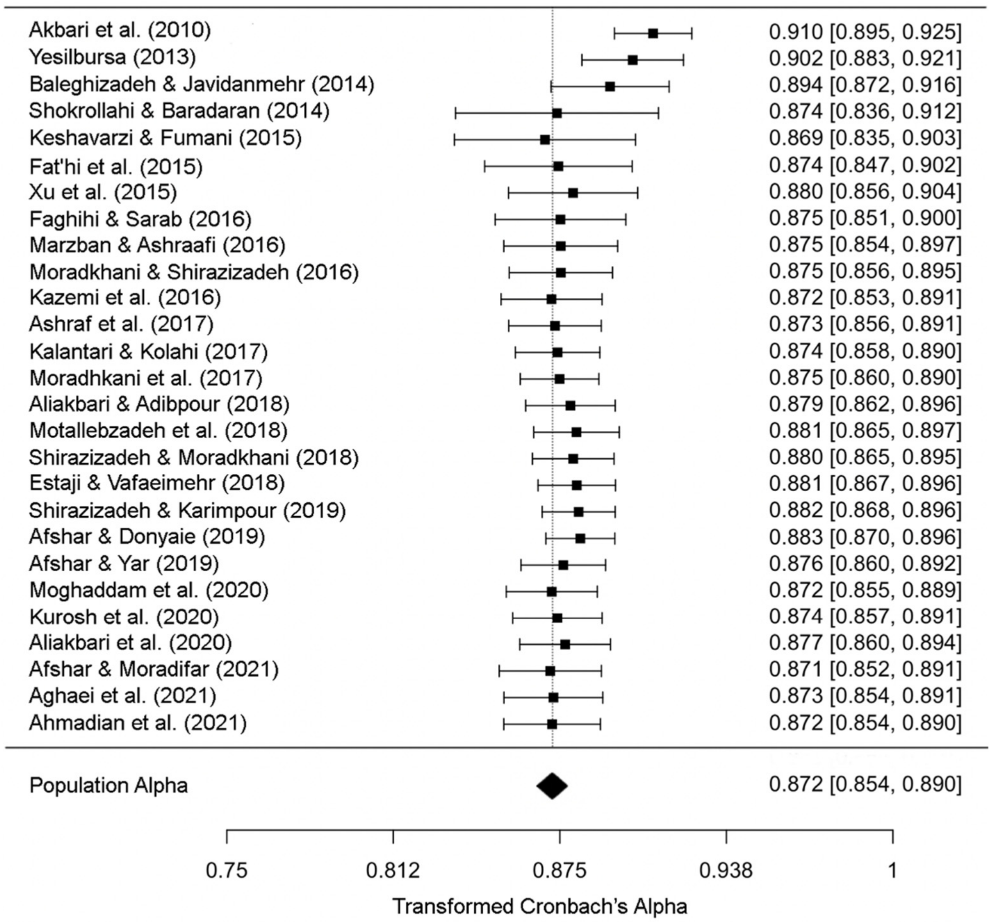

Figure 1. Forest Plot of Population and Samples of Transformed Alpha of ELTRI

4.2 Background Variable Analysis

Table 2. Meta-Regression Results Using Mixed-Effect Model

Moderator variables	b	SE	z	95% C.I		p-value
				Lower	Upper	
Intercept	1.039	0.105	9.921	0.834	1.245	< 0.001
Article type	0.234	0.025	9.359	0.185	0.283	< 0.001
Percent female	-0.023	0.140	-0.161	-0.297	0.252	0.872
Mean score on scale	0.001	0.001	0.069	-0.002	0.002	0.945
Mean of teaching experience	-0.016	0.003	-4.788	-0.022	-0.009	< 0.001

Note: b = regression coefficients, SE = standard error of estimate, z = z-value, where < 1.96 indicates significance

In this study, the factors used as moderator variables were article type, percentage of female subjects, mean scores on ELTRI, and the mean teaching experience of the samples in each study. The results of the meta-regression analysis are presented in Table 1.

The results of the meta-regression analysis using the mixed-effect model produced a non-significant Q-value of 0.2195, with a p-value of 0.639, by adding four moderator variables. This indicates a larger within-subjects variation within a study, rather than variation across studies. Thus, the addition of moderators was justified for explaining the heterogeneity found in the random-effect model.

Furthermore, based on Table 1, the standard regression coefficient (beta) shows that two of the four moderator variables have a significant effect on the high and low alpha values of the 27 ELTRI articles. First, article type has a significant positive impact on Cronbach's alpha ($b = 0.234$, $p < 0.001$). This can be interpreted as showing that the articles that focus on testing the psychometric characteristics of ELTRI as a measurement instrument (coded as 1) tend to have higher reliability than those focusing on the use of ELTRI for substantive research. Second, the mean of the teaching experience of the samples from each study has a significant negative impact on Cronbach's alpha values ($b = -0.016$, $p < 0.001$). This means that a study in which participants have a lower teaching experience will show higher reliability.

5. Discussions

This study provides an overview of the procedure for implementing alpha coefficient meta-analysis, a new method that researchers can use to estimate the population alpha coefficient from a number of reliability studies using the same instrument. In simple terms, after the data are collected and ready to be processed, the alpha coefficient meta-analysis procedure begins with the transformation of the alpha coefficient from each study and weighting it according to the number of samples and items used in each study. The next step is to calculate an alpha estimate for the population (weighted mean transformed alpha) and calculate the confidence interval. Alpha for a population is an estimated value that describes the Cronbach's alpha magnitude of a measuring instrument in the population.

After calculating the alpha estimates for the population, the next step is to perform the alpha homogeneity test on the study sample. If the homogeneity test results show that the data from the study sample are homogeneous, no further analysis is required. However, if the sample data are heterogeneous, further analysis is necessary to determine the factors that contribute to the heterogeneity of alpha data. Statistical methods such as meta-regression can be used to explore the factors that influence data heterogeneity.

This study provides an example of how the meta-analysis method is used to synthesize the alpha coefficients statistically from several study samples. Although the computational process for implementing alpha coefficient meta-analysis is quite complex, there are now many software tools available that can help researchers apply this kind of analysis.

For researchers in applied linguistics and social sciences, the alpha coefficient meta-analysis method can provide practical benefits for choosing a measurement tool to be used for research (Plonsky & Derrick, 2016). Given that information on the alpha coefficient of an instrument used in a study can be easily obtained, researchers can collect this information even without understanding complex statistical methods (Hedges & Vevea, 1998). Additionally, applied linguistics researchers can use the meta-analysis method, an advanced psychometric method, to quantitatively synthesize various research results with similar objectives and statistical tests to obtain more convincing conclusions (Hedges & Olkin, 1985).

Furthemore, considering that the issue of high and low reliability between studies is often examined from the perspective of test characteristics—such as the number of items and response format (e.g., Voss et al., 2000)—alpha meta-analysis provides another point of view for estimating population alpha, which is the RG of a measuring instrument, and examines the factors that influence population heterogeneity (e.g., Hedges & Vevea, 1998). Considering the use of ELTRI with original reliability as reported by Akbari et al. of 0.91 (e.g., Shahivand & Moradkhani, 2020), which only applies to the sample in their study, the population alpha estimate given by this study (0.872), which is the RG of several ELTRI studies, can be used as a reference (e.g., Vassar & Bradley, 2010) for the use of ELTRI.

Regarding the finding of moderator variables that significantly affect alpha coefficients, article-type variables (e.g., 1 = substantive; 0 = measurement) were often previously included in various alpha meta-analysis studies (e.g., Lane et al., 2002; Vassar & Crosby, 2008). In this study, articles that focused on testing the psychometric characteristics of the ELTRI tended to have higher reliability. This finding is reasonable considering that articles of this type tend to use advanced data analysis methods that require large samples with good sampling designs and perform psychometric analysis of item quality so that only the best items are included in the analysis (e.g., Akbari, 2010; Xu et al. al., 2015).

Regarding the significant effect of teaching experience, many articles use ELTRI and provide information about the teaching experience. However, to the best of our knowledge, this is the first study to include the aforementioned variable in the alpha meta-analysis. This study found that the reliability of ELTRI tends to be high in samples with low teaching experience. From a theoretical perspective, studies using ELTRI often include comparisons between novice and experienced EFL teachers (Kalantari & Kolahi, 2017) and have found a relationship between reflective teaching as measured by ELTRI and "the continuity of experience" (Kurosh et al., 2020). The findings of the present study strengthen the rationale for considering a sample group based on teaching experience. Substantive researchers in the field of ELT and applied linguistics can further explore these findings.

ELTRI is one of the most popular instruments in the ELT field. Thus, other and newer instruments in the ELT field, such as the English Language Teacher Learning Scale (Aliakbari & Malmir, 2017), Technological Pedagogical Content Knowledge for EFL (Bostanciioglu & Handley, 2018), and Teacher Classroom Leadership Scale in the EFL context (Khany & Ghasemi, 2021), can be analyzed using the same method. Considering that the use of an instrument in a study is often preceded by a literature review and the availability of information on the reliability coefficient of the instrument, as well as the easy accessibility of software to perform a meta-analysis of alpha coefficients, the present study constitutes a bridge for applied researchers in applying meta-analysis of alpha coefficients.

The question of how many articles are required as a minimum sample for an alpha meta-analysis study often arises when conducting a meta-analysis. The answer may not be straightforward, considering that the availability of data for each instrument is different. For example, Bonnett (2010) performed a meta-analysis with only nine study samples. Therefore, researchers can further explore the limitations of meta-analysis in the context of a small sample and its impact on the estimation results (e.g., Hedges & Vevea, 1996; Lin, 2018).

Finally, the following is a brief description of the alpha coefficient meta-analysis procedure that researchers can carry out: (1) determine the measuring instrument to be used; (2) formulate an operational definition of the study population; (3) specify inclusion and exclusion criteria for the study sample; (4) search for articles containing study reports on measuring instruments; (5) extract and encode data by collecting alpha coefficient values and sample sizes from each study, as well as other relevant moderator variables; (6) enter data into the program (e.g., Metafor package on R); (7) perform the computation of meta-analysis of alpha coefficient, homogeneity test, and analysis of factors that affect the variability of the alpha coefficient; (8) create a forest plot containing a graphical representation of the meta-analysis of alpha coefficients; and (9) draw conclusions from the meta-analysis of the alpha coefficient.

6. Conclusions

The results of this study indicate that ELTRI has an excellent population-level reliability coefficient ($\hat{\rho} = 0.872$). These results show that the ELTRI is an instrument that can generate accurate scores when used to assess English language teaching practices. The implication of this study suggests that ELTRI can be used in future studies to evaluate reflective teaching practice for teachers teaching English as a second language (L2)—which encompasses practical, cognitive, metacognitive, affective, critical, and moral reflection— and to investigate factors influencing teaching practices. The methodological implication of the findings of this study on the meta-analysis of coefficient alpha is that the population-level reliability of ELTRI, or its reliability generalization, can be used as a reference for use instead of the original version of ELTRI (see Akbari et al., 2010), which was based only on that study sample.

References

*Afshar, H. S., & Donyaie, S. (2019). English teachers' perception of critical pedagogy: any discrepancy between perception and actual classroom implementation? *Applied Research on English Language, 8*(1), 79-114.

*Afshar, H. S., & Moradifar, M. (2021). The structural interplay between critical cultural awareness, institutional identity, self-efficacy, reflective teaching, and job performance of EFL teachers. *International Journal of Society, Culture & Language, 9*(1), 14-29.

*Afshar, H. S., & Yar, S. H. (2019). Investigating the relationship between attitude towards professional development, reflective teaching, self-efficacy, and job performance of Iranian english as a foreign language teachers. *Teaching English Language, 13*(2), 147-179.

*Aghaei, P., Bavali, M., & Behjat, F. (2021). Relationship between professional identity and reflectivity: a case of Iranian EFL teachers. *International Journal of Foreign Language Teaching & Research, 9*(34), 11-23.

*Ahmadian, M., Maftoon, P., & Birjandi, P. (2021). A mixed-method study of teacher diary as a reflective instrument for the english language teacher development. *Journal of Modern Research in English Language Studies, 8*(2), 1-24. https://doi.org/10.30479/jmrels.2020.12002.1495

Ahrens, R. D. B., Lirani, L. D. S., & Francisco, A. C. D. (2020). Construct validity and reliability of the Work Environment Assessment Instrument WE-10. *International Journal of Environmental Research and Public Health, 17(*20), 7364. https://doi.org/10.3390/ijerph17207364

*Akbari, R., Behzadpoor, F., & Dadvand, B. (2010). Development of english language teaching reflection inventory. *System, 38*, 211-227. https://doi.org/10.1016/j.system.2010.03.003

*Aliakbari, M., & Adibpour, M. (2018). Reflective EFL education in Iran: existing situation and teachers' perceived fundamental challenges. *Eurasian Journal of Educational Research, 77*, 129-144.

Aliakbari, M., & Malmir, B. (2017). Development and validaation of an English language teacher learning scale. *Cogent Education, 4*(1), 1292613. https://doi.org/10.1080/2331186X.2017.1292613

*Aliakbari, M., Khany, R., Adibpour, M. (2020). EFL teachers' reflective practice, job satisfaction, and school context variables: Exploring possible relationships. *TESOL Journal, 11*:461. https://doi.org/10.1002/tesj.461

*Ashraf, H., Samir, A., & Yazdi, M. T. (2017). Brain dominance quadrants and reflective teaching among ELT teachers: A relationship study. *International Journal of English Linguistics, 7*(2), 63-72.

*Baleghizadeh, S., & Javidanmehr, Z. (2014). Exploring EFL teachers' reflectivity and their sense of self-efficacy. *e-International Journal of Educational Research, 5*(3), 19-38.

Biber, D., Nekrasova, T., & Horn, B. (2011). *The effectiveness of feedback for L1-English and L2-writing development: a meta-analysis' TOEFL iBT* (Research report No.14). Educational Testing Services.

Bonett, D. G. (2010). Varying coefficient meta-analytic methods for Alpha reliability. *Psychological Methods, 15*(4), 368-385. https://doi.org/10.1037/a0020142

Bostancioglu, A., & Handley, Z. (2018). Developing and validating a questionnaire for evaluating the EFL 'Total PACKage': Technological Pedagogical Content Knowledge (TPACK) for English as a Foreign Language (EFL). *Computer Assisted Language Learning, 31*(5-6), 572-598. https://doi.org/10.1080/09588221.2017.1422524

Botella, J., Suero, M., & Gambara, H. (2010). Psychometric inferences from a meta-analysis of reliability and internal consistency coefficients. *Psychological Methods, 15*(4), 386-397. https://doi.org/10.1037/a0019626

Brannick, M. T. & Zhang, N. (2013). Bayesian meta-analysis of coefficient Alpha. *Research Synthesis Methods, 4*, 198-207.

Chalmers, R. P. (2018). On misconceptions and limited usefulness of Ordinal Alpha. *Educational and Psychological Measurement, 78*(6), 1056-1071. https://doi.org/10.1177/0013164417727036

Chapelle, C. A. (2012). *The encyclopedia of applied linguistics*. Blackwell Publishing Ltd.

Cheung, M. W.-L. (2015). *Meta-analysis: a structural equation modeling approach*. John Wiley & Sons, Ltd

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

Cronbach, L. J. (1951). Coefficient Alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297-334. https://doi.org/10.1007/BF02310555

Derrick, D. J. (2016). Instrument reporting practices in second language research. *TESOL Quarterly, 50*(1), 132-153. https://doi.org/10.1002/tesq.217

Educational Testing Service. (2020). *Reliability and comparability of TOEFL iBT scores* (Vol. 3). Educational Testing Service (ETS).

*Estaji, M., & Vafaeimehr, R. (2018). Teacher's reflection and its components as predictors of EFL teacher's sense of classroom management. *Iranian Journal of Applied Language Studies, 10*(2), 27-58.

*Faghihi, G., & Sarab, M. R. A. (2016). Teachers as reflective practicioners: a survey on iranian english teachers' reflective practice. *The Journal of Teaaching Language Skills, 7*(4), 57-86.

*Fat'hi, J., Ghaslani, R., & Parsa, K. (2015). The relationship between post-method pedagogy and teacher reflection: A case of Iranian EFL teachers. *Journal of Applied Linguistics and Language Research, 2*(4), 305-321.

Gagen, T. (2019). *The predictive validity of IELTS scores: a meta-analysis*. Master's degree thesis, The University of Western Ontario

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher, 5*(10), 3–8. https://doi.org/10.3102/0013189X005010003

Greco, L. M., O'Boyle, E. H., Cockburn, B. S., & Yuan, Z. (2018). Meta-analysis of coefficient Alpha: A reliability generalization study. *Journal of Management Studies, 55*(4), 583-618. https://doi.org/10.1111/joms.12328

Haktanir, A., Kurnaz, M. F., & Gökalp, Z. S. (2024). Reliability generalization meta-analysis of the Brief Self-Control Scale (BSCS): Reliability evidence across age groups and languages. *Measurement and Evaluation in Counseling and Development*. Advance online publication. https://doi.org/10.1080/07481756.2023.2301289

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Academic Press.

Hedges, L. V., & Vevea, J. L. (1996). Estimating effect size under publication bias: small sample properties and robustness of random effects selection model. *Journal of Educational and Behavioral Statistics, 21*(4), 299-332. https://doi.org/10.3102/10769986021004299

Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods, 3*(4), 486–504. https://doi.org/10.1037/1082-989X.3.4.486

Henson, R. K., Kogan, L. R., & Vacha-Haase, T. (2001). A reliability generalization study of the teacher efficacy scale and related instruments. *Educational and Psychological Measurement, 61*(3), 404-420. https://doi.org/10.1177/00131640121971284

Higgins, J. P. T., & Thompson, S. G. (2002). Quantifiying heterogeneity in a meta-analysis. *Statistics in Medicine, 21*(11), 1539-1558. https://doi.org/10.1002/sim.1186

Hoekstra, R., Vugteveen, J., Warrens, M. J., & Kruyen, P. M. (2019). An empirical analysis of alleged misunderstandings of coefficient alpha. *International Journal of Social Research Methodology, 22*(4), 351-364. https://doi.org/10.1080/13645579.2018.1547523

Holland, D. F., Kraha, A., Zientek, L. R., Nimon, K., Fulmore, J. A., Johnson, U. Y., Ponce, H. F., Aguilar, M. G., & Henson, R. K. (2018). Reliability generalization of the Motivated Strategies for Learning Questionnaire: a meta-analytic view of reliability estimates. *SAGE Open, 8*(3). https://doi.org/10.1177/2158244018802334

*Kalantari, S., & Kolahi, S. (2017). The relationship between novice and experienced EFL teachers' reflective teaching and their burnout. *Journal of Professional Capital and Community, 2*(3), 169-187. https://doi.org/10.1108/JPCC-12-2016-0032

*Kazemi, A., Bazregarzadeh, Z., & Firoozi, M. (2016). Teacher reflectivity revisited: is teaching reflectivity gaining a foothold in Iran? *Journal of Language Teaching and Research, 7*(5), 993-1001. http://dx.doi.org/10.17507/jltr.0705.22

*Kesharvazi, S., & Fumani, M. R. F. Q. (2015). The impact of teachers' reflectivity and gender on their intellectual excitement and interpersonal teaching style. *Theory and Practice in Language Studies, 5*(3), 525-534.

Khany, R., & Ghasemi, F. (2021). Development and validation of teacher classroom leadership scale in EFL context. *International Journal of Leadership in Education, 24(*4), 513-532. https://doi.org/10.1080/13603124.2019.1591516

*Kurosh, S., Yousefi, M. H., & Kashef, S. H. (2020). Iranian teachers' reflective teaching practice in relation to self-efficacy perceptions: investigating teachers' discipline. *Reflective Practice*, https://doi.org/10.1080/14623943.2020.1748591

Lane, G. G., White, A. E., & Henson, R. K. (2002). Expanding reliability generalization methods with KR-21 estimates: An RG study of the Coopersmith Self-esteem Inventory. *Educational and Psychological Measurement, 62*(4), 685-711. https://doi.org/10.1177/0013164402062004010

Liang, J., Shou, Y., Wang, M.-C., Deng, J., & Luo, J. (2021). Alabama Parenting Questionnaire-9: A reliability generalization meta-analysis. *Psychological Assessment, 33*(10), 940–951. https://doi.org/10.1037/pas0001031

Lin, L. (2018). Bias caused by sampling error in meta-analysis with small sample sizes. *PloS ONE, 13*(9), e0204056. https://doi.org/10.1371/journal.pone.0204056

Loewen, S., & Plonsky, L. (2015). *An A-Z of applied linguistics research methods*. Springer.

*Marzban, A., & Ashraafi, N. (2016). Assessing reflective thinking skills in EFL/ESL instructors based on differences in graduation degree and nationality. *Reflective Practice*, http://dx.doi.org/10.1080/14623943.2016.1206879

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods, 23*(3), 412-433. https://doi.org/10.1037/met0000144

Miller, B. K., Nicols, K. M., Clark, S., Daniels, A., & Grant, W. (2018). Meta-analysis of coefficient Alpha for scores on the Narcissistic Personality Inventory. *PLoS ONE, 13*(12): e0208331. https://doi.org/10.1371/journal.pone.0208331

*Moghaddam, R. G., Davoudi, M., Adel, S. M. R., & Amirian, S. M. R. (2019). Reflective teaching through journal writing: a study on EFL teachers' reflection-for-action, reflection-in-action, and reflection-on-action. *English Teaching & Learning*, https://doi.org/10.1007/s42321-019-00041-2

*Moradkhani, S., & Shirazizadeh, M. (2016). Context-based variations in EFL teachers' reflection: the case of public schools versus private institutes in Iran. *Reflective Practice*, http://dx.doi.org/10.1080/14623943.2016.1267002

*Moradkhani, S., Raygan, A., & Moein, M. S. (2017). Iranian EFL teachers' reflective practices and self-efficacy: Exploring possible relationships. *System, 65*, 1-14. https://doi.org/10.1016/j.system.2016.12.011

*Motallebzadeh, K., Ahmadi, F., & Hosseinnia, M. (2018). The relationship between EFL teachers' reflective practices and their teaching effectiveness: A structural equation modeling approach. *Cogent Psychology, 5*: 1424682. https://doi.org/10.1080/23311908.2018.1424682

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). McGraw-Hill.

Okada, K. (2015). Bayesian meta-analysis of cronbach's coefficient Alpha to evaluate informative hypotheses. *Research Synthesis Methods, 6*, 333-346.

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

Peterson, R. A. (1994). A meta-analysis of cronbach's coefficient Alpha. *Journal of Consumer Research, 21*(2), 381-391. https://doi.org/10.1086/209405

Plonsky, L., & Derrick, D. J. (2016). A meta-analysis of reliability coefficients in second language research. *The Modern Language Journal, 100*(2), 538-553. https://doi.org/10.1111/modl.12335

Rahayu, W., Putra, M. D. K., Faturochman., Meiliasari., Sulaeman, E., & Koul, R. B. (2022). Development and validation of Online Classroom Learning Environment Inventory (OCLEI): the case of Indonesia during the COVID-19 pandemic. *Learning Environments Research, 25*(1), 97-113. https://doi.org/10.1007/s10984-021-09352-3

Raykov, T., & Marcoulides, G. (2010). *Introduction to psychometric theory*. Taylor & Francis.

Raykov, T., & Marcoulides, G. A. (2019). Thanks coefficient alpha, we still need you! *Educational and Psychological Meassurement, 79*(1), 200-210. https://doi.org/10.1177/0013164417725127

Rodriguez, M. C., & Maeda, Y. (2006). Meta-analysis of coefficient Alpha. *Psychological Methods, 11*(3), 306-322. https://doi.org/10.1037/1082-989X.11.3.306

Sen, S. (2021). A reliability generalization meta-analysis of Runco Ideational Behavior Scale. *Creativity Research Journal*. https://doi.org/10.1080/10400419.2021.1960719

Shahivand, E. S., & Moradkhani, S. (2020). The relationship between EFL teachers' trait emotional intelligence and reflective practices: a structural equation modeling approach. *Innovation in Language Learning and Teaching, 14*(5), 466-480. https://doi.org/10.1080/17501229.2019.1620241

*Shirazizadeh, M., & Karimpour, M. (2019). An investigation of the relationships among EFL teachers' perfectionism, reflection and burnout. *Cogent Education, 6*, 1667708. https://doi.org/10.1080/2331186X.2019.1667708

*Shirazizadeh, M., & Moradkhani, S. (2018). Minimizing hurnout through reflection: the rocky road ahead of EFL teachers. *Teaching English Language, 12*(1), 135-154.

*Shokrollahi, M., & Baradaran, A. (2014). On the relationship between Iranian EFL teachers' perfectionism and their reflectivity. *International Journal of Language Learning and Applied Linguistics World, 7*(3), 13-28.

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

Soetjipto, H. P., Putra, M. D. K., Widhiarso, W., & Khakim, Z. (2023). Assessment of the psychometric properties of the Indonesian version of the General Self-Efficacy Scale-12 (GSES-12) in a sample of Indonesian high school students. *Psikohumaniora: Jurnal Penelitian Psikologi*, 8(2), 259-274. https://doi.org/10.21580/pjpp.v8i2.18024

Streiner, D. L., & Norman, G. R. (1989). *Health measurement scales: a practical guide to their development and use* (2nd ed.). Oxford University Press.

Suryadi, B., Hayat, B., & Putra, M. D. K. (2020). Evaluating psychometric properties of the Muslim Daily Religiosity Assessment Scale (MUDRAS) in Indonesian samples using the Rasch model. *Mental Health, Religion and Culture, 23*(3-4), 331-346. https://doi.org/10.1080/13674676.2020.1795822

Vacha-Haase, T. (1998). Reliability generalization: exploring variance in measurement error affecting score reliability. *Educational and Psychological Measurement, 58*(1), 6–20. https://doi.org/10.1177/0013164498058001002

Vassar, M., & Bradley, G. (2010). A reliability generalization study of coefficient alpha for the Life Orientation Test. *Journal of Personality Assessment, 92*(4), 362-370. https://doi.org/10.1080/00223891.2010.482016

Vassar, M. & Bradley, G. (2011). A reliability generalization meta-analysis of coefficient Alpha for the Reynolds Adolescent Depression Scale. *Clinical Child Psychology and Psychiatry, 17*(4), 519-527. https://doi.org/10.1177/1359104511424998

Vassar, M., & Crosby, J. W. (2008). A reliability generalization study of coefficient alpha for the UCLA Loneliness Scale. *Journal of Personality Assessment, 90*(6), 601-607. https://doi.org/10.1080/00223890802388624

Vicent, M., Rubio-Aparicio, M., Sánchez-Meca, J., & Gonzálvez, C. (2019). A reliability generalization meta-analysis of the Child and Adolescent Perfectionism Scale. *Journal of Affective Disorders, 245,* 533-544. https://doi.org/10.1016/j.jad.2018.11.049

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36*(3), 1-48. https://doi.org/10.18637/jss.v036.i03

Voss, K. E., Stem, D. E., & Fotopoulos, S. (2000). A comment on the relationship between coefficient alpha and scale characteristics. *Marketing Letters, 11*(2), 177-191. https://doi.org/10.1023/A:1008146924781

Watanabe, Y., & Koyama, D. (2008). A meta-analysis of second language cloze testing research. *Second Language Studies, 26*(2), 103-133.

Wilkinson, L., & APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*(8), 594-604. https://doi.org/10.1037/0003-066X.54.8.594

*Xu, J., Li, B., & Curtis, A. (2015). Validating an English Language Teaching Reflection Inventory in a Chinese EFL context. *System, 49*, 50-60. https://doi.org/10.1016/j.system.2014.10.014

*Yesilbursa, A. (2013). Construct validation of the english language teaching reflective inventory with a sample of Turkish university EFL instructors. *English Language Teaching, 6*(5), 28-37. http://dx.doi.org/10.5539/elt.v6n5p28

Yörük, S., & Sen, S. (2022). A reliability generalization meta-analysis of the creative achievement questionnaire. *Creativity Research Journal*. Advance online publication. https://doi.org/10.1080/10400419.2022.2148073

Zumbo, B. D., Gadermann, A. M., & Zeisser, C. (2007). Ordinal versions of coefficients alpha and theta for Likert rating scales. *Journal of Modern Applied Statistical Methods, 6*, 21-29.

Acknowledgments

Not applicable.

Authors contributions

Not applicable.

Funding

Not applicable.

Competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Informed consent

Not applicable.

Ethics approval

Not applicable.

Provenance and peer review

Not commissioned; externally double-blind peer reviewed.

Data availability statement

The data that support the findings of this study are available on request from the corresponding author.

Data sharing statement

No additional data are available.