Jose Fabian Elizondo-Gonzalez[1]
Universidad de Costa Rica

Original scientific paper
UDC: 37.022
http://doi.org/10.5937/IstrPed2502471F

==========

OPTIMIZING INTER-RATER RELIABILITY IN FOREIGN LANGUAGE CONSTRUCTED-RESPONSE ASSESSMENTS

Abstract: This study examines inter-rater reliability in a constructed-response English proficiency test developed by the Foreign Language Assessment Program (PELEx) in Costa Rica. Thirty university instructors completed three writing tasks aligned with A2, B2, and C1 CEFR bands, each scored by two trained raters. Inter-rater reliability was estimated using percent agreement, Cohen's weighted kappa, intraclass correlation coefficients (ICCs), and Generalizability Theory (G-Theory). While traditional estimators suggested good to excellent reliability, G-Theory revealed additional sources of error not accounted for by kappa or ICC, particularly prompt-related variability. For example, in the C1 task, person-by-prompt interaction accounted for over 20% of total variance. These findings suggest that while rater training remains important, prompt-related variability must also be addressed to ensure fairness and score comparability. Incorporating calibrated prompts and structured scoring protocols across proficiency levels may strengthen the reliability of constructed-response tasks, especially in high-stakes settings.

Keywords: constructed-response assessment, Costa Rica, EFL, language assessment, reliability

Introduction

Constructed-response[2] assessment deals with values assigned to observable outcomes using measurement tools, such as rubrics or checklists (Scott, 2003). Unlike traditional norm-referenced assessments, constructed-response assessments are criterion-referenced, comparing test-takers' performance against predetermined standards rather than to other test-takers. However, for such assessments to be meaningful and equitable, they must not only align with construct-relevant criteria but also provide empirical evidence of their reliability and validity (Burger & Burger, 2005; Wagner, 2020).

Reliability is crucial in constructed-response assessment, as test-taker behaviors are rated by judges, something that can introduce multiple sources of error in the scoring process. Reliability, or the consistency/precision in test scores (Revelle & Condon, 2019), can be compromised when judges are not properly trained or calibrated for specific scoring tasks, not adhering to scoring protocols or being too lenient or severe (Hallgren, 2012; Lyness et al., 2021). Also, raters can suffer from a 'halo' effect, where test-taker performance is rated based on impressions and not established protocols, introducing judgement bias (Landy & Sigall, 1974; Wetzel et al., 1981). What is more, rater fatigue can affect the reliability of scoring procedures, in that it has been shown that test-taker performance is more severely rated for the first tasks in a testing batch, and last tasks are more leniently rated (Mahshanian & Shahnazari, 2020; Wagner, 2020). These sources of error variance pose a serious threat to the construct validity of constructed-response assessments, as they may distort score interpretations and ultimately undermine test fairness.

[1] josefabian.elizondo@ucr.ac.cr; https://orcid.org/0000-0003-4819-0213
[2] In this study, the terms "constructed-response" and "performance-based" will be considered interchangeable.

Measuring Reliability in Constructed- response Assessment

Though there are multiple methods to account for test reliability, the most-commonly used techniques have been designed to gather evidence of the reliability of the measure itself in selected-response tests (e.g., with multiple-choice formats) (Tavakol & Dennick, 2011). This type of reliability is estimated using observed test-taker responses through mathematical modeling, instead of human raters' judgement. Some of these include Cronbach's alpha (Cronbach, 1951) in Classical Test Theory (CTT) and McDonald's Omega (McDonald, 1999) in Structural Equation Modeling (SEM). However, these coefficients are not designed to be implemented when computing reliability in constructed-response assessment, for their purpose is not to analyze rater behavior but rather the consistency of the test or items themselves (Flemenbaum & Zimmermann, 1973).

Other methods have been established to estimate reliability in constructed-response assessment, including estimations for inter-rater reliability (IRR) and intra-rater reliability. IRR can be defined as the degree of agreement between two raters in assigning the same score to some variables (Hallgren, 2012; McHugh, 2012), while intra-rater reliability refers to raters' self-consistency while scoring performance-based tasks (Gwet, 2008). In its simplest form, IRR can be calculated as the proportion of agreements among raters relative to the total number of ratings. This is often referred to as percent agreement, and it is computed as:

$$IRR = \frac{\text{Number of agreements}}{\text{Total number of ratings}}$$

However, percent agreement alone is not considered a robust measure of IRR, because it does not account for agreement occurring by chance (Hallgren, 2012). More sophisticated methods, such as Cohen's kappa (for two raters) and intraclass correlation coefficients (ICCs) (for multiple raters), adjust for chance agreement and provide a more reliable estimate of rater consistency.

Cohen's kappa was designed for estimating agreement for nominal scales, whose results indicate the "proportion of agreement after chance agreement is removed from consideration" (Cohen, 1960, p. 40). Though useful, Cohen's kappa has some limitations. First, regarding research design, it assumes a fully crossed design, where all the raters provide a score to all observations; for example, if 10 essays need to be scored, and there are three raters hired for the scoring task, all three raters will provide a score for each of the 10 essays. Second, in terms of the numbers of raters, Cohen's kappa is employed for when the methodology includes two raters only (Hallgren, 2012), which may limit the possibilities in exploring rater behavior.

Similarly to Cohen's kappa, intraclass correlation coefficients (ICCs) is another frequently-used statistic to determine IRR. Unlike Cohen's kappa, it is suitable for interval and ratio variables, with designs including more than two raters and does not require fully-crossed designs, meaning not all raters need to score all responses. While ICC estimates the degree of agreement among raters, it also captures the magnitude of rater disagreement, making it a more flexible measure for performance-based assessments (Hallgren, 2012).

ICC is based on variance decomposition, meaning it evaluates how much of the total variability in scores is due to actual differences between test-takers rather than inconsistencies among raters. Different ICC models account for variability in rater-mediated scoring, and the choice depends on how raters are assigned to responses (Shrout & Fleiss, 1979). A one-way random effects model (Case 1) is appropriate when each response is rated by a different set of randomly assigned raters, such as in a speaking test where each individual is graded by a different examiner. A two-way random effects model (Case 2) is used when the same group of raters scores all responses, treating raters as randomly selected from a larger population. Finally, a two-way mixed effects model (Case

3) is suitable when the same fixed set of raters scores all responses, as when tasks are judged by a selected expert panel, not randomly chosen.

Though ICC is often preferred over Cohen's kappa for its flexibility, both methods share two key limitations. First, while they estimate reliability by measuring the consistency of ratings across judges, they do not fully account for the multiple sources of error variance inherent in constructed-response (and performance-based) assessments (Hallgren, 2012). These include variability in the tasks being scored, differences across test administrations, and topic-specific effects, all of which can influence rater judgments. Second, neither method provides a framework for optimizing reliability under different conditions. That is, Cohen's kappa and ICC estimate reliability based on the observed data but cannot model alternative scenarios—such as estimating reliability if the number of raters were changed or if additional sources of variance were explicitly incorporated into the analysis.

To address these limitations, Generalizability Theory (G-theory) was developed as a more comprehensive framework by decomposing error variance into multiple facets (e.g., raters, tasks, and test administrations), allowing for a more precise estimation of reliability (Burger & Burger, 2005; Shavelson & Webb, 1981). Unlike traditional agreement indices, which provide only a single reliability estimate based on observed conditions, G-theory allows researchers to systematically evaluate how different sources of measurement error affect score consistency and determine the most reliable testing conditions. This is achieved through two key types of analyses: Generalizability (G) studies and Decision (D) studies.

G-studies estimate the variance components associated with different facets, such as raters, tasks, or test occasions, identifying how much each contributes to overall score variability. These findings then inform D-studies, which focus on optimizing measurement designs by adjusting the number of raters, tasks, or test occasions to improve reliability, minimizing error, while maintaining feasibility (Webb & Shavelson, 2005). By modeling measurement conditions under different scenarios, G-theory not only quantifies reliability but also provides actionable insights for enhancing scoring procedures.

Reliability in Test Development for Foreign Language Proficiency Assessments

Reporting rater reliability is essential in any constructed-response assessment, but it becomes even more critical when validating a new test intended for large-scale, high-stakes use (De Champlain et al., 2016; Lynes et al., 2021; Norris & Lee, 2023). In such contexts, reliability is not just a technical concern but a key component of the test's construct validity argument to support test score interpretations (AERA, APA, & NCME, 2014). Establishing transparent, evidence-based scoring protocols is necessary to ensure that test-taker performance is measured consistently, particularly when human raters are involved. Without such studies, the dependability, or reliability, of rater-mediated scoring remains uncertain, limiting the interpretability of test results.

Because of the high impact foreign language proficiency assessments can have, such as determining college admissions or job promotions (ETS, 2020c), test developers attempting to validate newly created or high-stakes measures including constructed-response or performance-based tasks are also expected to provide reliability evidence of the methods employed to ensure accuracy in the scoring of said tasks. Examples of how this practice is intended to be met include the reliability evidence presented by some of the most well-known language proficiency tests available today: the Test of English as Foreign Language (TOEFL) iBT, TOEFL Essentials, and the Duolingo English Test (DET).

As a precursor of TOEFL Essentials, the reliability of writing tasks for the TOEFL iBT is estimated using Cronbach's alpha and test-retest methods, where tasks are scored by both human raters and

automated 'e-raters' (ETS, 2020b; ETS, 2020c). These go through constant calibration and training to ensure dependable scoring; for example, by analyzing writing responses of repeater samples, or by including 'monitor papers'—previously scored tasks— in new test batches for raters to score to measure expected rater performance (ETS, 2020a; ETS, 2020c). To estimate rater reliability, ETS reports using rater agreement rates that include exact and adjacent agreement (one point difference), and rater scores on each task are compared with the mean score for said task across raters (ETS, 2020c). Though this information is reported, it is unclear if the method employed is Cohen's kappa, intraclass correlation coefficients, G-theory, or a different method. Also, ETS does not seem to provide information about their inter-rater reliability coefficient practices; for example, reporting their minimum agreement level among raters or what their average inter-rater reliability agreement level is.

The newly created TOEFL Essentials has been designed as a multi-stage adaptive test (MST), to measure both academic and general English proficiency (ETS, 2021; Norris & Lee, 2023; Papageorgiou et al., 2021; Papageorgiou et al., 2022). In terms of its rater reliability in constructed-response tasks, research information is not readily available as it is for the TOEFL iBT, where an entire TOEFL Research Insight Series is available to stakeholders. However, Papageorgiou et al. (2021) report that to ensure scoring quality, TOELF Essential writing tasks go through 'similar' processes as TOEFL iBT tasks, though the differences are not explicitly communicated. No mention of how reliability coefficients are estimated—beyond briefly mentioning 'rater agreement'— is reported either. Therefore, it seems that the inter-rater reliability evidence of this new test, which has a different construct and operationalization overall, is dependent on the TOEFL iBT research evidence.

Similar to the tasks in TOEFL Essentials, Duolingo English Test (DET) writing tasks are computer adaptive and are scored by an automated algorithm created by machine learning (ML) and natural language processing (NLP) experts (DET, 2021a; DET, 2021b; Settles et al., 2020); hence, DET constructed-response tasks are not scored by human raters but by trained machines. Using data from a 2018 sample, DET (2021a) indicates test developers employed Cohen's kappa as their index of inter-rater reliability and reported that the IRR among human raters and machine raters was greater— at $\kappa = 0.79$— than the one observed among human raters, $\kappa = 0.77$. Though the report called "Analysis of the Scoring and Reliability for the Duolingo English Test" (DET, 2021b) allegedly included a more detailed section on IRR of constructed responses, that portion seems to have been removed from the publicly available version of the document. These newer studies and explicit reports on multiple sources of validity and reliability evidence seem to stem from research that has referred to DET as having a "black box nature of procedures [that] lack independent empirical, peer-reviewed research validating the use of DET scores" (Wagner, 2020, p. 15).

Purpose

Having this context in mind, the Foreign Language Assessment Program (PELEx, for its acronym in Spanish) has created a foreign language proficiency test for English (Prueba de Dominio Lingüístico), composed of four subtests, measuring reception (reading and listening comprehension) and production activities (speaking and writing) based on the Common European Framework of Reference for Languages (CEFR; Council of Europe, 2020; Araya et al., 2022; Céspedes & Araya, 2023). Though internal protocols have been created to guarantee accuracy in assessing the constructed-response tasks of the test using percent agreement to estimate IRR, no empirical studies have been conducted to determine the most appropriate method for estimating rater reliability or to explore how scoring consistency can be optimized.

For that reason, this study aims at examining the reliability of rater scoring in PELEx's constructed-response assessments by comparing four estimation methods: percent agreement, Cohen's kappa, intraclass correlation coefficients (ICCs), and G-Theory. Additionally, it investigates how a Decision

(D)-study within the G-theory framework can enhance scoring reliability across CEFR proficiency bands, through the following research questions:

1. How do Cohen's kappa, ICCs, and G-Theory compare in estimating the reliability of written task scores across CEFR bands?
2. How can a Decision (D)-study in Generalizability Theory optimize scoring reliability for written tasks in constructed-response assessments at each CEFR band?

Methodology

Sample

The sample consists of 30 university instructors at a Costa Rican university, who took PELEx's English proficiency test to seek language certification for promotion purposes within their departments in March of 2022. The de-identified dataset for this study was provided by PELEx; hence, it was not subjected to IRB approval. The dataset only included test-takers' scores per CEFR band in each task per rater. No demographic information was collected nor provided, such as gender, age, or ethnicity.

Measure

Proficiency in PELEx's written production subtest is understood as the ability to write texts written in non-technical English in both formal and informal contexts at the regional and global levels in the personal and educational domains as developed by CEFR. Some of the skills to be tested range from writing "simple phrases and sentences about themselves and imaginary people" to writing "clear, well-structured expositions of complex subjects, emphasizing the relevant salient issues" (Council of Europe, 2020).

The subtest consists of three tasks where test-takers are to demonstrate proficiency at the A2, B2, and C1 levels, accordingly, based on the CEFR. For A2 and B2 tasks, individuals are given a unique prompt from an item bank, while for C1 items test-takers are given two prompt options to choose from. Some of the tasks include formats such as replying to a post-card, writing an email, or defending an opinion in a structured essay. Through a fully-crossed design, each task is rated by two experts who are applied linguists in the field of Teaching English as a Foreign Language (TEFL). They are to provide a score using a holistic rubric designed by PELEx, which includes communicative language activities, strategies, and linguistic competences, such as grammatical accuracy and vocabulary control, for each CEFR band (Council of Europe, 2020). In total, six different raters are employed in each test scoring protocol: two for A2 tasks, two for B2 tasks and two for C1 tasks. The raters provide their scores independently, and those are sent to one test administrator who compiles all the scores across raters for each task. In case of finding absolute rater disagreement, a third rater—usually a senior PELEx rater— is employed as a tiebreaker. Table 1 shows how tasks are scored and coded for this study.

Table 1. *Writing task scoring and coding in ascending order*

CEFR Band	Performance descriptor	Coding
A2 tasks		
NA	Does not meet expectations	0
A1	Almost meets expectations at A2 level but fails to do so	1
A2 or A2+	Meets or exceeds expectations at A2 level	2
B2 tasks		
NA	Does not meet expectations	0
B1	Almost meets expectations at B2 level but fails to do so	1
B2 or B2+	Meets or exceeds expectations at B2 level	2
C1 tasks		

NA	Does not meet expectations	0
B2	Almost meets expectations at C1 level but fails to do so	1
C1	Meets expectations at C1 level	2

Reliability Estimators

To answer the research questions, raw percent agreement, Cohen's kappa, intraclass correlation coefficients (ICCs), and Generalizability Theory (G-Theory) were used to assess the reliability for each task (A2, B2, and C1 CEFR bands).

Cohen's kappa was estimated as:

$$\kappa = \frac{p_0 - p_c}{1 - p_c}$$

where p_0 is proportion of agreement across judges and p_c is the proportion of agreement expected by chance (Cohen, 1960). As a rule of thumb, κ statistics are typically reported in ranges indicating agreement strength: poor (< 0.00) slight (0.0 to 0.2), fair (0.21 to 0.40), moderate (0.41 to 0.60), substantial (0.61 to 0.80), and almost perfect (.81 to 1.0) (Landis & Kock, 1977). For this study, a variation of Cohen's kappa called Weighted kappa κ_w was employed to give partial credit to adjacent agreements using quadratic weights, and not depend solely on exact agreement (Cohen, 1968).

The ICC equations follow different models depending on how raters are assigned to responses. Shrout & Fleiss (1979) define these cases using the following linear models:

For Case 1 (One-Way Random Effects Model):

$$x_{ij} = \mu + b_j + w_{ij}$$

where:

- μ is the overall mean score across all responses and judges.
- b_j represents the true score component of response j, capturing actual response quality.
- w_{ij} includes all remaining sources of variability, including judge differences, response-judge interactions, and measurement error.

For Case 2 and Case 3 (Two-Way Models):

$$x_{ij} = \mu + a_i + b_j + (ab)_{ij} + e_{ij}$$

- μ is the grand mean score across all responses and judges,
- a_i represents the effect of judge i, accounting for whether they tend to score higher or lower overall (strictness/leniency effect),
- b_j is the true score component of response j, representing how well the response performed,
- $(ab)_{ij}$ is the interaction effect, capturing whether judge i scores response j higher or lower than expected based on their usual scoring tendency and the response's general quality, and
- e_{ij} represents remaining random error, including momentary inconsistencies in judging.

Because of the fully-crossed design of this study with fixed, specialized raters, the equation for Case 3, a two-way mixed effects model, was employed. The results obtained from ICCs are interpreted as IRR levels, where < 0.40 indicates poor IRR, 0.40 to 0.59 fair IRR, 0.60 to 0.74 good IRR, and > 0.75 excellent IRR (Cicchetti, 1994).

A Generalizability Study (G-study) was conducted using an ANOVA-based model for tasks at the A2 and B2 levels. All test takers responded to the same prompt, and each response was scored independently by two raters. This setup allowed for a fully crossed design with two facets: person and rater. The G-study estimated variance components attributable to persons (examinees), raters, and residual error, using the following model:

$$X_{pi} = \mu + (\mu_p - \mu) + (\mu_i - \mu) + e_{pi}$$

where:

- μ is the grand mean (average true scores across examinees),
- μ_p is the true score for examinee p,
- $\mu_p - \mu$ is the examinee effect, and
- $\mu_i - \mu$ is the rater effect.
- e_{pi} is the residual.

This model assumes that all persons were rated by all raters under the same task conditions—an assumption met in the A2 and B2 datasets.

For the C1 task, however, the design differed. Test-takers were given a choice between two writing prompts, and each responded to only one. As a result, prompt and person were not fully crossed: it was not possible to independently estimate a main effect for prompts. Instead, prompt-related variability was modeled as part of a person-by-prompt interaction, which was treated as a distinct source of error in the G-study. The model for the C1 task was therefore modified to reflect this structure:

$$X_{pi} = \mu + \phi_p + (\phi\lambda)_p + e_{pi}$$

where:

- ϕ_p is the person effect,
- $(\phi\lambda)_p$ represents the person-by-prompt interaction, and
- e_{pi} is the residual (including rater-related error and other unmodeled sources).

In both designs, the G-study variance estimates were used to conduct Decision Studies (D-studies), which projected generalizability coefficients (ρ) under different rater conditions (e.g., 2, 5, or 10 raters). The generalizability coefficient was calculated as:

$$\rho = \frac{\sigma_p^2}{\sigma_p^2 + \frac{\sigma_e^2}{n}}$$

where σ_p^2 is the true-score variance and σ_e^2 is the error variance divided by the number of raters n.

The results from the D-studies were interpreted using the same agreement levels proposed for ICCs by Cicchetti (1994), where values below 0.40 are considered poor, 0.40– 0.59 fair, 0.60– 0.74 good, and 0.75 or above excellent.

Results

For the A2 task, both raters assigned identical scores to all participants, indicating perfect agreement. As a result, reliability coefficients such as Cohen's kappa, intraclass correlation coefficient (ICC), and generalizability coefficients could not be meaningfully calculated, as these methods require variability between raters. Given this perfect agreement, inter-rater reliability at the A2 level can be considered maximal (i.e., 1).

Therefore, the following analyses focus on the B2 and C1 tasks, where variability in rater scores allowed for meaningful estimation. Table 2 summarizes the reliability results for B2 tasks.

Table 2. *Reliability coefficients for B2 tasks using multiple estimation methods*

Estimator	Coefficient	Statistic	*p*	Rater reliability
Percent agreement	0.73	-	-	Good
Cohen's Weighted kappa	0.25	*z*= 2.07	0.04	Fair
Intraclass correlations c.	0.48	*F*= 1.91	0.04	Fair
D-study (G-theory) 2 raters	0.48	-	-	Fair
D-study (G-theory) 5 raters	0.70	-	-	Good
D-study (G-theory) 7 raters	0.76	-	-	Excellent

The raw percent agreement, which was used by PELEx to report rater consistency in this sample, indicated good agreement and suggests some degree of scoring stability. However, both Cohen's weighted kappa and ICC yielded statistically significant but only fair levels of agreement. Generalizability Theory analyses further indicated that reliability under the current design is fair. D-study projections suggest that at least seven raters would be required for the B2 task to reach excellent levels of inter-rater reliability.

As for C1 tasks, Table 3 summarizes the results obtained using multiple IRR estimation methods.

Table 3. *Reliability coefficients for C1 tasks using multiple estimation methods*

Estimator	Coefficient	Statistic	*p*	Rater reliability
Percent agreement	0.70	-	-	Good
Cohen's Weighted kappa	0.63	*z*= 3.63	< 0.001	Substantial
Intraclass correlations c.	0.81	*F*=5.13	< 0.001	Excellent
D-study (G-theory) 2 raters	0.59	-	-	Fair
D-study (G-theory) 5 raters	0.78	-	-	Excellent

Percent agreement suggested good consistency, while Cohen's weighted kappa indicated substantial agreement after adjusting for chance. The ICC further pointed to excellent reliability between raters. However, these estimators focus solely on rater-related error and do not account for additional sources of measurement error - most notably, differences introduced by the choice of prompts. Since test-takers selected between two writing prompts, any variability related to prompt difficulty or scoring consistency across prompts is not captured by traditional reliability measures.

Generalizability Theory was applied to account for these multiple sources of error. Variance component analysis revealed that 42% of the total variance was due to true differences between test-takers, while 21% was attributed to the person-by-prompt interaction and 37% to residual (rater-related) error. This pattern suggests that a substantial portion of the error variance stems from prompt-related differences. For instance, while both prompts were designed to assess C1-level writing, 64% of responses to Prompt 2 reached the C1 standard, compared to just 40% for Prompt 1 - putting those who selected Prompt 1 at a disadvantage.

D-studies showed that with the current two-rater design, the generalizability coefficient was 0.59, reflecting fair reliability, while increasing the number of raters to five would substantially improve reliability estimates.

Discussion

Inter-rater reliability (IRR) was examined in the constructed-response writing tasks included in the PELEx English proficiency test at the A2, B2, and C1 CEFR levels using multiple estimation methods.

The A2 tasks achieved perfect IRR, representing the most dependable results within this subtest. In the context of language certification, such a high level of agreement in scoring constructed-response tasks reflects an ideal scenario - one in which test-takers can trust that their performance is being evaluated fairly and consistently.

Though historically used at PELEx, percent agreement may fall short in capturing the complex sources of variability involved in scoring constructed-response tasks (Hallgren, 2012). For the B2 writing tasks, estimators that adjust for chance agreement - such as Cohen's weighted kappa and the ICC - suggest that IRR remains at a fair level. These findings point to the potential value of enhanced rater training to improve consistency. Similarly, results from D-studies indicate that achieving excellent reliability under current conditions would require at least seven raters, a scenario that is both impractical and financially burdensome. To address these challenges, PELEx may benefit from implementing structured rater calibration and certification protocols, such as those used by ETS and aim to achieve higher levels of inter-rater reliability like those reported for DET (ETS, 2020a; ETS, 2020c; DET, 2021a). Doing so would strengthen scoring consistency and transparency, particularly in high-stakes assessment contexts.

The C1 task findings raise important considerations about the design and scoring of constructed-response items in proficiency assessments. While traditional estimators such as kappa and ICC suggested relatively strong agreement between raters, the G-theory analysis revealed that rater agreement alone does not fully explain score variability. A notable proportion of error stemmed from person-by-prompt interactions, indicating that test-takers may have been differentially advantaged or disadvantaged depending on the prompt they chose. This may be explained by the nature of the prompts and their alignment with test-takers' backgrounds. Prompt 2, which focused on higher education funding, likely resonated more with the sample of college instructors, who are professionally engaged with this topic. In contrast, Prompt 1, on gun violence, may have required a different type of discourse or posed greater emotional or linguistic challenges. These differences suggest that prompt content can interact with examinee characteristics in ways that influence performance, reinforcing the need for careful prompt selection and validation in proficiency testing.

This is particularly relevant in high-stakes testing, where such variability can undermine perceptions of fairness and comparability. In this context, improving rater consistency - while still important - will have limited impact unless prompt design and scoring procedures are also addressed. To strengthen score reliability and fairness, test developers may consider building a bank of calibrated prompts at each proficiency level - tasks that have been pre-evaluated for equivalence in difficulty and scoring behavior. Such a system would allow for more consistent assessment across administrations and provide a foundation for monitoring and improving reliability over time using G-Theory or similar frameworks. In sum, the results point to the need for more robust prompt equivalence procedures and potentially rethinking whether prompt choice should be offered at all when score comparability is a priority.

It should be noted that these findings should be interpreted with caution. As Atilgan (2013) suggests, G-theory analyses generally require at least 50 participants for robust estimation. Given the current sample size, G-theory reliability estimates may be unstable. PELEx is therefore encouraged to replicate this study with a larger sample to confirm and refine these initial findings.

Conclusion

This study demonstrates how different reliability estimation methods can yield substantially different conclusions about inter-rater agreement in constructed-response language proficiency assessments. While traditional methods such as weighted kappa and ICC suggested fair to excellent reliability, Generalizability Theory revealed additional sources of error that these methods do not capture - particularly those related to prompt choice. The findings from the C1 task underscore how

prompt-related variability can structurally limit the reliability of scores, even under ideal rater conditions. To enhance the consistency and fairness of scoring, test developers may benefit from building calibrated prompt banks and implementing rater certification protocols. These practices can help ensure that performance-based assessments not only meet technical standards of reliability but also support valid and equitable interpretations of test scores across CEFR proficiency bands.

References:

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

Araya Garita, W., Elizondo González, J., & González Ramírez, A. (2022). Getting stakeholders acquainted with the rationale behind the construct of the English language proficiency test of the University of Costa Rica for the Ministry of Education of Costa Rica. *Estudios de Lingüística Aplicada, 75*, 119–143. https://doi.org/10.22201/enallt.01852647p.2022.75.1013

Atilgan, H. (2013). Sample size for estimation of G and Phi coefficients in generalizability theory. *Eurasian Journal of Educational Research, 51*, 215–227.

Burger, S. E., & Burger, D. L. (1994). Determining the validity of performance-based assessment. *Educational Measurement: Issues and Practice, 13*(1), 9–15. https://doi.org/10.1111/j.1745-3992.1994.tb00779.x

Céspedes Araya, J., & Araya Garita, W. (2023). Standardized evaluation experience at Universidad de Costa Rica: PELEx and AI integration. *UCIENCIA 2023*. https://repositorio.uci.cu/jspui/handle/123456789/10768

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6*(4), 284–290. https://doi.org/10.1037/1040-3590.6.4.284

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin, 70*(4), 213–220. https://doi.org/10.1037/h0026256

Council of Europe. (2020). *Multilateralism 2020: Annual report of the Secretary General of the Council of Europe (Annual Activity Report 2020)*. Council of Europe. https://rm.coe.int/annual-report-sg-2020/1680a05193

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

De Champlain, A. F., Gotzmann, A., & Qin, S. (2016). Assessing the reliability of performance assessment scores: Some considerations in selecting an appropriate framework. *Journal of Graduate Medical Education, 8*(4), 504–506. https://doi.org/10.4300/JGME-D-15-00751.1

Duolingo English Test (DET). (2021a). *Analysis of the validity, design and development of the Duolingo English Test*. https://dy8n3onijof8f.cloudfront.net/media/resources/standards/validity.pdf

Duolingo English Test (DET). (2021b). *Analysis of the scoring and reliability for the Duolingo English Test*. https://dy8n3onijof8f.cloudfront.net/media/resources/standards/scoring.pdf

Educational Testing Service (ETS). (2020a). *TOEFL iBT test framework and test development*. https://www.ets.org/pdfs/toefl/toefl-ibt-insight-s1v1.pdf

Educational Testing Service (ETS). (2020b). *TOEFL research*. https://www.ets.org/pdfs/toefl/toefl-ibt-insight-s1v2.pdf

Educational Testing Service (ETS). (2020c). *Reliability and comparability of TOEFL iBT scores*. https://www.ets.org/pdfs/toefl/toefl-ibt-insight-s1v3.pdf

Educational Testing Service (ETS). (2021). *Why choose the TOEFL Essentials test?* https://www.ets.org/toefl/test-takers/essentials/about/why.html

Flemenbaum, A., & Zimmermann, R. L. (1973). Inter- and intra-rater reliability of the Brief Psychiatric Rating Scale. *Psychological Reports, 33*(3), 783–792. https://doi.org/10.2466/pr0.1973.33.3.783

Gwet, K. L. (2008). Intrarater reliability. In *Wiley encyclopedia of clinical trials* (Vol. 4, No. 2, pp. 473–485). Wiley.

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology, 8*(1), 23–34. https://doi.org/10.20982/tqmp.08.1.p023

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174. https://doi.org/10.2307/2529310

Landy, D., & Sigall, H. (1974). Beauty is talent: Task evaluation as a function of the performer's physical attractiveness. *Journal of Personality and Social Psychology, 29*(3), 299–304.

Lyness, S. A., Peterson, K., & Yates, K. (2021). Low inter-rater reliability of a high stakes performance assessment of teacher candidates. *Education Sciences, 11*(10), 648. https://doi.org/10.3390/educsci11100648

McDonald, R. P. (1999). *Test theory: A unified treatment*. Lawrence Erlbaum Associates.

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica, 22*(3), 276–282. https://doi.org/10.11613/BM.2012.031

Mahshanian, A., & Shahnazari, M. (2020). The effect of raters' fatigue on scoring EFL writing tasks. *Indonesian Journal of Applied Linguistics, 10*(1), 1–13

Norris, J. M., & Lee, J. (2023). *The effectiveness of the TOEFL® Essentials™ test for distinguishing English proficiency levels* (Research Memorandum No. RM-23-07). Educational Testing Service. https://www.ets.org/Media/Research/pdf/RM-23-07.pdf

Papageorgiou, S., Davis, L., Norris, J. M., Garcia Gomez, P., Manna, V. F., & Monfils, L. (2021). *Design framework for the "TOEFL® Essentials"™ test 2021* (Research Memorandum No. RM-21-03). Educational Testing Service. https://files.eric.ed.gov/fulltext/ED617531.pdf

Papageorgiou, S., Davis, L., Ohta, R., & Garcia Gomez, P. (2022). Mapping TOEFL® Essentials™ test scores to the Canadian Language Benchmarks. *ETS Research Report Series, 2022*(1), 1–42.

Revelle, W., & Condon, D. M. (2019). Reliability from α to ω: A tutorial. *Psychological Assessment, 31*(12), 1395–1411. https://doi.org/10.1037/pas0000754

Scott, S. (2003). Practical implications of reliability and performance-based assessments. *General Music Today, 16*(3), 18–22.

Settles, B., LaFlair, G. T., & Hagiwara, M. (2020). Machine learning–driven language assessment. *Transactions of the Association for Computational Linguistics, 8*, 247–263. https://doi.org/10.1162/tacl_a_00310

Shavelson, R. J., & Webb, N. M. (1981). Generalizability theory: 1973–1980. *British Journal of Mathematical and Statistical Psychology, 34*(2), 133–166. https://doi.org/10.1111/j.2044-8317.1981.tb00580.x

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420–428. https://doi.org/10.1037/0033-2909.86.2.420

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

Wagner, E. (2020). Duolingo English Test, revised version July 2019. *Language Assessment Quarterly, 17*(3), 300–315. https://doi.org/10.1080/15434303.2020.1771343

Webb, N. M., & Shavelson, R. J. (2005). Generalizability theory: Overview. In *Encyclopedia of statistics in behavioral science* (Vol. 2, pp. 717–719). Wiley.

Wetzel, C. G., Wilson, T. D., & Kort, J. (1981). The halo effect revisited: Forewarned is not forearmed. *Journal of Experimental Social Psychology, 17*(4), 427–439. https://doi.org/10.1016/0022-1031(81)90045-2

Biographical notes:

Jose Fabian Elizondo Gonzalez is an EFL instructor and researcher who works at Universidad de Costa Rica. He holds two master's degrees, one in Education Administration and a second one in Teaching English as a Foreign Language. Currently, he is a PhD candidate in the Educational Psychology Department at the University of Kansas, USA.