Tabaran Institute of Higher Education
*International Journal of Language Testing*
Vol. 16, No. 1; 2026, March

ISSN 2476-5880
https://www.ijlt.ir

# Examining Measurement Invariance of a C-Test Across Gender Using Multiple-Group Item Response Theory

Salokhitdinova Navruza[1]*, Hamroyev Gulom[2], Temirova Matluba[3], Xolmatova Ziroat[4], Khudayberganov Khudaybergan[5], Matkarimov Inomjon[6], I. B. Sapaev[7], Van Truong Chu[8]

**ARTICLE INFO**

**Article History:**
Received: June 2025
Accepted: July 2025

**KEYWORDS**
C-Test
Genfer DIF
Invariance
MG-IRT
PCM

**ABSTRACT**

Examination of measurement invariance is essential for cross-group comparisons. This study investigates the measurement invariance of a C-Test across gender using a Multiple-Group Item Response Theory framework. A C-Test, composed of six passages totaling 120 items, was administered to 256 intermediate-level English as a Second Language (ESL) learners at Termez University in Uzbekistan. To assess whether the test functioned equivalently for male and female participants with similar language proficiency, MG-IRT modeling was employed using the Partial Credit Model. Model fit indices, item-level statistics, and differential item functioning (DIF) were evaluated through Root Mean Square Difference (RMSD) and Mean Difference (MD) values. Results indicated high reliability (α = .96, EAP = .93), good model fit, and ordered item thresholds. Importantly, none of the six C-Test passages demonstrated meaningful gender-based DIF, suggesting that the instrument exhibits strong measurement invariance. These findings support the validity and fairness of the C-Test in assessing general language proficiency across gender and highlight the utility of MG-IRT models in language test validation.

[1] PhD, Associate Professor, Termiz University of Economics and Service, 190111 Termez, Uzbekistan. https://orcid.org/0009-0000-2227-5812
[2] PhD, Samarkand State Medical University. Samarkand, Uzbekistan. https://orcid.org/0009-0002-8855-9919
[3] Teacher, Department of Primary Education, Termez State Pedagogical Institute, I.Karimov street 288b, Termez, Surxondaryo, Uzbekistan. https://orcid.org/0009-0001-8340-2501
[4] DSc, Associate professor, Department of Sciences and training management,
Kokand State University, Kokand, Uzbekistan. https://orcid.org/0009-0008-5031-4063
[5] Urgench State University, 14, Kh.Alimdjan str, Urganch, Khorezm, Uzbekistan. https://orcid.org/0009-0003-5484-5471
[6] Associate professor, Department of Economy, Mamun University, Uzbekistan. https://orcid.org/0000-0002-6783-8591
[7] 1. Head of the Department Physics and Chemistry, Tashkent Institute of Irrigation and Agricultural Mechanization Engineers, National Research University, Tashkent, Uzbekistan.
2. Scientific researcher of the University of Tashkent for Applied Science, Tashkent, Uzbekistan.
3. School of Engineering, Central Asian University, Tashkent, Uzbekistan.
4. Western Caspian University, Scientific researcher, Baku, Azerbaijan.
5. Baku Eurasian University, Baku, AZ 1073, Azerbaijan. https://orcid.org/0000-0003-2365-1554
[8] PhD student in Management, Centre for Postgraduate Studies, Swiss Information and Management Institute (SIMI Swiss) & Asia Metropolitan University (AMU), 63000 Cyberjaya, Selangor, Malaysia. https://orcid.org/0009-0009-4843-9868

## 1. Introduction

Language assessment plays a critical role in educational settings, guiding decisions about placement, instruction, and achievement. Among the many formats for assessing general language proficiency, the C-Test has gained recognition for its practical efficiency and theoretical robustness. Originally developed by Raatz and Klein-Braley (1981), the C-Test is a form of reduced redundancy testing that builds upon the classical cloze procedure. In a typical C-Test, the second half of every second word is deleted in a series of short, coherent texts, and the test-taker is required to restore the missing portions. This design leverages linguistic redundancy and contextual cues, requiring the integration of lexical, grammatical, and discourse knowledge. Because of these properties, C-Tests are believed to assess general language competence more holistically than traditional discrete-point items (Grotjahn, 2010).

Beyond its theoretical appeal, the C-Test has demonstrated strong psychometric properties. Studies have reported high reliability coefficients, reduced susceptibility to test-taking strategies, and ease of scoring, making it a favorable instrument in both research and educational practice (Baghaei & Tabatabaee, 2015; Hassan et al., 2024). Furthermore, the C-Test has shown validity evidence across a variety of contexts including university placement, entrance examinations, and large-scale language proficiency studies (Dörnyei & Katona, 1992). However, the growing usage of the C-Test also underscores the necessity of examining its fairness across diverse subpopulations—particularly gender-based comparability—to ensure its equitable application in educational decisions.

The concept of measurement invariance (MI) lies at the heart of fair assessment. MI is the statistical property indicating that a test measures the same latent trait across different groups (Meredith, 1993). In educational measurement, the absence of MI may suggest differential item functioning (DIF), whereby certain items systematically favor one group over another despite equal underlying ability. For example, if male and female examinees with the same language proficiency levels have different probabilities of correctly completing a C-Test item, the item would exhibit gender-based DIF. Such violations undermine the validity of group comparisons and pose serious ethical and interpretive concerns (Van de Vijver & Tanzer, 2004; Zumbo, 2007).

Traditionally, researchers relied on Classical Test Theory (CTT) to examine item bias, but these methods are often sample-dependent and limited in their ability to model the complexities of test performance. In contrast, Item Response Theory (IRT) offers a powerful psychometric framework for modeling the interaction between item characteristics and latent ability, allowing for parameter-level comparisons across groups (Korompot et al., 2024). Within IRT, Multiple-Group IRT (MG-IRT) models extend the framework to explicitly test measurement invariance by allowing or constraining item parameters (difficulty, discrimination) across groups. MG-IRT is particularly suited to investigating uniform DIF, i.e., constant bias across ability levels and non-uniform DIF where bias varies by ability (Bock & Zimowski, 1997; Thissen et al., 1993).

Despite the strengths of IRT, few studies have applied MG-IRT models to examine the C-Test specifically in the context of gender-based measurement invariance. This represents a critical gap in the literature, given the extensive use of C-Tests in both high-stakes testing and classroom assessment.

A limited but growing number of studies have touched upon this topic. For instance, Baghaei (2010) conducted an analysis using Rasch modeling and found evidence that certain C-Test items exhibited misfit and DIF, suggesting that they did not perform equivalently across subgroups. Forthmann et al. (2020) employed various IRT models to examine speeded C-Tests and concluded that model selection had significant implications for DIF detection, particularly across gender and academic majors. Similarly, Alpizar et al. (2023) proposed a multidimensional IRT approach to assess

C-Test passages and noted the importance of evaluating gender-based measurement bias to ensure the validity of test scores.

More recently, Schnoor et al. (2023) analyzed longitudinal invariance of C-Tests using structural equation modeling (SEM), revealing moderate to strong support for measurement stability but also identifying subtle variations across demographic groups. In their study, gender-related bias was not the primary focus, but findings suggested the potential for differential functioning that warrants deeper investigation.

Other studies have explored related constructs using similar methods. For example, Lowe (2015) examined the test anxiety inventory using MG-IRT and demonstrated how gender-based DIF could distort observed scores. Although not directly applied to C-Tests, such research reinforces the transferability of MG-IRT techniques to language testing contexts. Similarly, Schooner et al. (2023) evaluated general language skill development using C-Tests and found their invariance across groups to be tenuous in certain settings, advocating for robust measurement models that account for demographic differences. Atmawinata et al. (2025) examined the psychometric properties of PIRLS 2021 in Kazakhstan. They examined gender DIF using MG-IRT and showed that most of the MD and RMSD values are within the acceptable range.

Another compelling argument for investigating gender DIF in C-Tests stems from sociolinguistic and test-taking behavior research, which suggests that males and females may engage with language tasks differently due to cognitive styles, educational backgrounds, or motivational factors (Breland et al., 1995; O'Loughlin, 2002). Without rigorous invariance testing, such group-based differences may confound score interpretations and potentially disadvantage one gender over another. This concern is particularly urgent in contexts where C-Test scores influence academic admissions or placement decisions.

The current study addresses this important issue by applying MG-IRT modeling to investigate whether a widely used C-Test exhibits measurement invariance across gender. Specifically, we seek to determine whether the test functions equivalently for male and female participants with comparable levels of language proficiency. This analysis includes the examination of DIF at the item level as well as invariance at the test structure level, allowing us to draw nuanced conclusions about the fairness of the instrument.

In addition to advancing methodological rigor, this study contributes to the evidence-based development of language assessments that are both psychometrically sound and socially just. As the field continues to emphasize test fairness and equity, particularly in multilingual and multicultural settings, our findings will inform not only C-Test developers but also educators, policymakers, and language testers more broadly.

## 2. Method

### *2.1 Participants*

The participants in this study were 256 undergraduate students enrolled in English as a Second Language (ESL) courses at Termez University, located in Uzbekistan. The sample comprised 55% female (n = 141) and 45% male (n = 115) students, reflecting the gender distribution within the university's language program. The mean age of the participants was 22.47 years with a standard deviation (SD) of 3.11, indicating a moderately homogeneous age distribution. All participants were intermediate-level English learners, enrolled in language proficiency development courses at the time of data collection. Participation was voluntary, and informed consent was obtained prior to administration of the test.

*2.2 Instruments*

The primary instrument used in this study was a teacher-developed C-Test, designed to measure general English language proficiency. The C-Test consisted of six independent passages, each approximately 80–100 words in length and of comparable difficulty. Every passage contained 20 gaps, resulting in a total of 120 C-Test items. The construction of the C-Test adhered to established guidelines (Raatz & Klein-Braley, 1981), where the second half of every second word (excluding the first and last sentences) was systematically deleted. This approach aimed to preserve coherence while enabling the assessment of integrative language skills such as morphology, syntax, and lexical knowledge.

Each gap required the test-taker to accurately restore the missing portion of the word. The first and last sentences of each passage remained intact to provide contextual anchoring. The C-Test was administered in a classroom setting under standardized conditions, and responses were scored dichotomously (correct/incorrect) at the item level but later aggregated into passage-level scores for analysis.

## 3. Analysis

To examine the presence of measurement invariance and potential differential item functioning (DIF) across gender, the data were analyzed using a MG-IRT framework. Specifically, each of the six passages was treated as a polytomous super-item, representing the sum of correct responses within that passage. This method allowed the modeling of partial credit data while circumventing the local dependence issue, as recommended in prior research on C-Test analysis (Baghaei & Christensen, 2023; Forthmann et al., 2020; Grotjahn, 2010).

MG-IRT modeling was conducted using the partial credit model (PCM, Masters, 1982), suitable for ordered categorical data. Model estimation was conducted in a single step. An MG-IRT was estimated with gender as a grouping variable. In this procedure, invariance analysis indicates whether the group-based ICCs (item characteristic curve) match the ICC obtained from the entire sample (Baghaei & Robitzsch, 2025). If the group ICCs are close to the whole-sample ICC, it is evidence that there is no DIF, and the whole-sample ICC explains the item within the groups. However, if the group-based ICCs do not match the overall ICC, it means that the item shows DIF and the overall ICC cannot explain the item behavior within the groups.

Evidence of significant differences between group and overall ICCs was provided with the Root Mean Square Difference (RMSD) and Mean Difference (MD). All analyses were performed using the TAM package (Robitzsch et al., 2022) in R (R Core Team, 2024).

*3.1 Results*

Table 1 shows the descriptive statistics for the groups. The table shows that females overall have outperformed males. They are also more homogeneous with smaller standard deviation and variance. Cronbach’s alpha reliability of the C-Test with six items was .96.

**Table 1**

*Descriptive Statistics for the Groups and the Whole Sample*

| | Male | Female | Whole Sample |
|---|---|---|---|
| Mean | 26.35 | 29.33 | 28.67 |
| Median | 30 | 33 | 32 |

| | | | |
|---|---|---|---|
| Mode | 35 | 34 | 35 |
| SD | 8.99 | 6.88 | 7.44 |
| Variance | 80.90 | 47.45 | 56.12 |
| Min. | 8 | 10 | 8 |
| Max. | 36 | 36 | 36 |
| Range | 28 | 26 | 28 |

An MG-PCM with gender as a grouping variable was run to examine the items. Table 2 shows the item measures and fit statistics for the six C-Test passages. As the table shows all the items fit the Rasch model with infit and outfit mean square values between .50 to 1.50 (Linacre, 2025). The EAP reliability based on the MGIRT estimation was .93. The thresholds (τ) are all ordered and reasonably distanced (Linacre, 2025). The MG-IRT mean ability of males was 0, and the mean ability of females was .92, indicating a higher achievement for females.

**Table 2**

*Item Measures, Thresholds, and Fit Statistics*

| Item | Measure | $\tau_1$ | $\tau_2$ | $\tau_3$ | $\tau_4$ | $\tau_5$ | Infit | Outfit |
|---|---|---|---|---|---|---|---|---|
| 1 | -.08 | -6.01 | -2.93 | -.87 | 2.39 | 7.42 | .96 | 1.01 |
| 2 | -1.29 | -5.42 | -2.78 | -.01 | 2.18 | 6.02 | .87 | .76 |
| 3 | 1.45 | -4.99 | -2.93 | -.23 | 1.78 | 6.38 | .86 | .97 |
| 4 | .51 | -6.73 | -2.92 | .56 | 2.74 | 6.35 | .79 | .87 |
| 5 | -.66 | -4.21 | -3.09 | .10 | 2.31 | 4.90 | .72 | .69 |
| 6 | .07 | -4.71 | -.66 | 1.60 | 3.07 | 5.74 | .89 | .74 |

Table 3 shows the RMSD and MD values for males and females for the six C-Test items. According to OECD (2017), RMSD values <.12 indicate absence of DIF, while for MD, values < -.12 and >.12 indicate invariance. As Table 2 shows, all the values are within the acceptable range. Therefore, the six C-Test items do not show gender DIF.

**Table 3**

*RMSD and MD Fit Values for Males and Females*

| | RSMD | | MD | |
|---|---|---|---|---|
| | Male | Female | Male | Female |
| Item 1 | .04 | .02 | 0.00 | 0.00 |
| Item 2 | .04 | .03 | 0.00 | 0.00 |
| Item 3 | .07 | .03 | 0.00 | 0.00 |
| Item 4 | .06 | .03 | 0.00 | 0.00 |
| Item 5 | .05 | .03 | -.01 | 0.00 |
| Item 6 | .07 | .04 | 0.00 | 0.00 |

A comparison between the single-group PCM and MG-PCM using likelihood ratio test showed that the MGPCM fits significantly better than the single-group PCM, $\chi^2$=11.84, *df*=2, *p*=.002. Neverthelss, DIF as shown by RDMSD and MD statistics, was very small and negligible.

## 4. Discussion

The present study aimed to evaluate the measurement invariance of a teacher-developed C-Test across gender using a Multiple-Group Item Response Theory framework. Given the widespread use of the C-Test in educational contexts—ranging from classroom placement to high-stakes proficiency testing—ensuring fairness in test interpretation across demographic subgroups is both a psychometric and ethical imperative.

Overall, findings from this study indicate that the six passages of the C-Test functioned equivalently for male and female students. The RMSD and MD values for all items were well within acceptable thresholds (OECD, 2017), suggesting the absence of gender-based Differential Item Functioning (DIF). Moreover, fit statistics for each item fell within the expected Rasch model range (0.50 to 1.50), and the ordered and well-spaced thresholds further confirmed the psychometric soundness of the instrument. While a likelihood ratio test revealed that the MG-PCM model provided a statistically better fit than the single-group PCM model, the actual differences in item functioning were minimal and did not constitute practical significance.

These results offer strong support for the measurement invariance of the C-Test across gender, adding to the growing body of research that endorses the test's fairness and reliability. This finding aligns with previous work (e.g., Alpizar et al., 2023; Forthmann et al., 2020) that demonstrated the utility of IRT-based approaches for evaluating invariance and supports the suitability of the C-Test for use in diverse educational populations.

Interestingly, despite females outperforming males on average—consistent with some prior studies on gender and language assessment performance (Breland et al., 1995; O'Loughlin, 2002)—this difference appears to reflect genuine variation in underlying language proficiency rather than measurement bias. The absence of DIF suggests that the test is not inherently favoring one gender, but rather accurately capturing latent language ability across groups. This distinction is crucial for maintaining the validity of inferences drawn from test scores and for informing fair placement or instructional decisions.

Several conclusions emerge from these findings. First, the results reinforce the C-Test's validity as a tool for assessing general language proficiency without introducing systematic gender bias. This has direct relevance for educators, test developers, and policy-makers aiming to ensure equity in educational assessment. Second, by employing a MG-IRT framework and treating each C-Test passage as a polytomous super-item, the study offers a replicable analytic approach that mitigates local dependence and models partial credit appropriately—issues that have historically complicated C-Test analysis. And finally, given the psychometric fairness of the C-Test across gender, its continued use in classroom and institutional decision-making appears justified. Our findings show that score differences between males and females may be attributed to contextual and instructional variables rather than test bias.

## 5. Conclusion

From a pedagogical perspective, the results offer several important takeaways for language educators, curriculum designers, and assessment developers. First, the confirmation of measurement invariance across gender allows teachers and program administrators to use C-Test results with greater confidence, knowing that male and female learners are being assessed equitably. This is especially important in ESL/EFL contexts where test scores may influence placement into instructional tracks, scholarship eligibility, or progression through language programs.

Second, the findings underscore the value of integrative assessment tools like the C-Test, which evaluate multiple language domains simultaneously. Since no gender-based DIF was detected, the results suggest that the C-Test format does not inadvertently favor test-taking strategies, linguistic styles, or cognitive processes that are more common among one gender than the other. This reinforces the pedagogical suitability of the C-Test in diverse classrooms where learners bring varied educational and sociocultural backgrounds.

Third, these findings encourage continued adoption of data-driven approaches to assessment validation. Instructors and institutions should not only consider reliability and content validity but also actively investigate fairness and bias—particularly in settings with gender imbalances or where assessment outcomes have high stakes. Embedding fairness analyses into routine test development and review cycles will help ensure assessments support inclusive educational environments.

Finally, this study suggests that language educators should feel reassured using C-Tests for formative and summative purposes without introducing unintended gender bias. However, it also prompts ongoing vigilance; while this particular C-Test demonstrated fairness, future test forms, especially those including culturally specific content or complex discourse genres, should also be scrutinized for invariance. Equitable assessment must remain a dynamic and evidence-based practice.

Despite these contributions, some limitations should be acknowledged. First, the sample was restricted to intermediate-level learners from a single university in Uzbekistan, which may limit generalizability. Future studies should examine measurement invariance across broader linguistic and cultural contexts, proficiency levels, and demographic variables such as socioeconomic status or academic discipline. Second, while this study focused on gender as a binary construct, future research could explore invariance across a more inclusive spectrum of gender identities, in line with current psychometric and ethical best practices.

Future studies should examine measurement invariance using confirmatory factor analysis techniques and compare the results with those of MG-IRT. Furthermore, methods of addressing non-invariance such as partial invariance and alignment optimization should be investigated (Asparouhov & Muthén, 2014; Ravand, 2024; Sandoval-Hernandez et al., 2025).

The current study provided robust evidence that the C-Test, as implemented here, exhibits measurement invariance across gender. This supports its use as a fair and reliable measure of language proficiency in mixed-gender educational settings. As assessment practices continue to evolve, integrating rigorous statistical techniques like MG-IRT will be essential in upholding the principles of equity and validity that underpin responsible educational testing.

**Acknowledgments (obligatory)**
We would like to thank the anonymous reviewers for their comments.

**Declaration of Conflicting Interests (obligatory)**
The authors declare that they have no conflict of interest

**Funding (obligatory)**
This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

**Declaration of AI-Generated Content**
We have made no use of any generative AI aids for writing this paper.

## References

Alpizar, D., Li, T., Norris, J. M., & Gu, L. (2023). Psychometric approaches to analyzing C-tests. *Language Testing, 40*(1), 107–132. https://doi.org/10.1177/02655322211062138

Asparouhov, T., & Muthén, B. (2014). Multiple-group factor analysis alignment. *Structural Equation Modeling: A Multidisciplinary Journal, 21*(4), 495–508. https://doi.org/10.1080/10705511.2014.919210

Atmawinata, M. R., Herwina, W., Bulan, S., Wikanengsih, W., Darheni, N. & Tashtemirova, G. (2025). Item response theory analysis of the Progress in International Reading Literacy Study (PIRLS) 2021 in Kazakhstan. *International Journal of Language Testing, 15*(1), 122–152. doi: 10.22034/ijlt.2024.456241.1343

Baghaei, P., & Christensen, K. B. (2023). Modelling local item dependence in C-Tests with the Loglinear Rasch Model. *Language Testing Journal, 40*(3), 820–827. https://doi.org/10.1177/02655322231155109

Baghaei, P., & Robitzsch, A. (2025). A tutorial on item response modeling with multiple groups using TAM. *Educational Methods & Psychometrics, 3,* 1–14. https://dx.doi.org/10.61186/emp.2025.1

Baghaei, P., & Tabatabaee, M. (2015). The C-Test: An integrative measure of crystallized intelligence. *Journal of Intelligence, 3*, 46–58. https://doi.org/10.3390/jintelligence3020046

Baghaei, P. (2010). An investigation of the invariance of Rasch item and person measures in a C-Test. In R. Grotjahn (Ed.). *Der C-Test: Beiträge aus der aktuellen Forschung/ The C-Test: Contributions from Current Research* (pp.100-112). Peter Lang.

Bock, R. D., & Zimowski, M. F. (1997). Multiple group IRT. In van der Linden, W. J. & Hambleton, R. K. (Eds.) *Handbook of modern item response theory* (pp. 433–448). Springer. https://doi.org/10.1007/978-1-4757-2691-6-25

Breland, H. M., Bridgeman, B., & Fowles, M. E. (1995). *Writing assessment: A review of research and practice*. ETS.

Dörnyei, Z., & Katona, L. (1992). Validation of the C-Test among Hungarian EFL learners. *Language Testing*, 9(3), 187–206. https://doi.org/10.1177/026553229200900205

Forthmann, B., Grotjahn, R., Doebler, P., & Baghaei, P. (2020). A comparison of different item response theory models for scaling speeded C-Tests. *Journal of Psychoeducational Assessment, 38*, 692–705. Doi: 10.1177/0734282919889262

Grotjahn, R. (2010). *Der C-Test: Theoretische Grundlagen und praktische Anwendungen*. Peter Lang.

Hassan, A. Y., Jaafar, E. A., Al-Rawe, M. F. A., & Abdullah, S. S. (2024). Validation of C-Test among Iraqi EFL university students. *International Journal of Language Testing, 14*(2), 151–161. doi: 10.22034/ijlt.2024.446036.1329

Korompot, C. A., Siregar, I., Khursanov, N. I., Abdullaev, D. & Mohamed, K. M (2024). Investigating gender DIF in the reading comprehension section of the B2 First Exam. *International Journal of Language Testing, 14*(2), 57–66. doi: 10.22034/ijlt.2023.421011.1301

Linacre, J. M. (2025). *Winsteps® Rasch measurement computer program User's Guide*. Version 5.10.1. Portland, Oregon: Winsteps.com

Lowe, P. A. (2015). Should test anxiety be measured differently for males and females? *Journal of Psychoeducational Assessment*, 33(4), 305–315. https://doi.org/10.1177/0734282914549428

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*(2), 149–174. https://doi.org/10.1007/BF02296272

Meredith, W. (1993). Measurement invariance, factor analysis, and factorial invariance. *Psychometrika*, 58(4), 525–543.

OECD (2017). *PISA 2015 technical report*. OECD Publishing.

O'Loughlin, K. (2002). The impact of gender in the IELTS test. *IELTS Research Reports*, 4, 71–89.

Raatz, U., & Klein-Braley, C. (1981). The C-Test – A modification of the cloze procedure. *Language Testing*, 1(2), 113–138. https://doi.org/10.1177/026553228100100202

Ravand, H. (2024). Assessing measurement invariance in a university entrance exam: A comparison of multigroup confirmatory factor analysis alignment method vs. multigroup item response theory. *Educational Methods & Psychometrics*, *2*, 1–20. https://dx.doi.org/10.61186/emp.2024.4

R Core Team (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Robitzsch, A., Kiefer, T., & Wu, M. (2022). *TAM: Test Analysis Modules*. R package version 4.1-4. https://CRAN.Rproject.org/package=TAM

Sandoval-Hernandez, A., Carasco, D., & Eryilmaz, N. (2025). Alignment optimization in International Large-Scale Assessments: A scoping review and future directions. *Educational Methods and Psychometrics*, *3*, 1–25. https://dx.doi.org/10.61186/emp.2025.3

Schnoor, B., Hartig, J., Klinger, T., Naumann, A., & Usanova, I. (2023). Measuring the development of general language skills in English as a foreign language—Longitudinal invariance of the C-test. *Language Testing*, *40*(3), 796–819. https://doi.org/10.1177/02655322231159829

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In Holland, P.W., & Wainer, H. (Eds.), *Differential Item Functioning* (pp. 67–113). Erlbaum.

Van de Vijver, F. J. R., & Tanzer, N. K. (2004). Bias and equivalence in cross-cultural assessment: An overview. *Revue Européenne de Psychologie Appliquée*, 54(2), 119–135.

Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly*, 4(2), 223–233. https://doi.org/10.1080/15434300701375832