Optimizing Reliability Scoring for ILSAs

Ji Yoon Jung **Ummugul Bezirhan** **Matthias von Davier**
TIMSS & PIRLS International Study Center at Boston College
{jiyoon.jung, bezirhan, vondavim}@bc.edu

Abstract

This study proposes an innovative method for evaluating cross-country scoring reliability (CCSR) in multilingual assessments, using hyperparameter optimization and a similarity-based weighted majority scoring within a single human scoring framework. Results show that this approach provides a cost-effective and comprehensive assessment of CCSR without the need for additional raters.

1 Introduction

Constructed response (CR) items are valued for their ability to assess students' higher order thinking skills, offering deeper insights into student performance compared to multiple choice items (Livingston, 2009; Scully, 2017). However, their widespread use in large-scale assessments has been constrained by concerns about human scoring reliability. While extensive rater training and structured scoring protocols can enhance inter-rater reliability, rater effects such as leniency, severity, and the halo effect often persist (Myford & Wolfe, 2003; Yamamoto et al., 2017).

These scoring challenges are particularly pronounced in international large-scale assessments (ILSAs). In multilingual contexts, achieving high consistency among human raters from diverse cultural and linguistic backgrounds is difficult, even with centralized scoring guides (Wang & Li, 2020). The substantial time, effort, and resources required for global human rater training, scoring vast numbers of responses, and monitoring scoring procedures across multiple countries further complicate the process.

Cross-country scoring reliability (CCSR), designed to measure international scoring consistency (von Davier et al., 2023) in the Progress in International Reading Literacy Study (PIRLS), exemplifies these challenges. This valuable measure operates as a separate, additional burden alongside the main scoring process and encounters significant logistical hurdles. It evaluates scoring consistency using a common set of 200 English language responses for specific PIRLS reading items, but its scope is critically limited to human raters who are either native English speakers or proficient in English. Consequently, the conventional CCSR approach assesses a narrow subset of responses and relies on an underrepresented rater pool. This restricts its ability to provide a comprehensive assessment of scoring consistency across the full range of CR items and participating countries.

To address these logistical and methodological limitations, we recently proposed a novel reliability scoring framework that combines similarity-based majority voting (Jung et al., under review).

The current study focuses on the systematic optimization of that framework through hyperparameter tuning while also providing a transparent step-by step implementation of the full pipeline. This method aims to offer a more efficient and reliable measure of cross-country scoring consistency, reducing dependency on extensive human rater resources.

2 Background

Human scoring in multilingual assessments presents significant challenges, primarily due to difficulties in maintaining consistency across different human raters, languages, and countries (Jung et al., 2025; Okubo et al., 2023). The inherent linguistic and sociocultural diversity among raters may influence the interpretation of student responses and the application of scoring guides, introducing systematic variance in scoring outcomes (Ercikan & Por, 2020; Wang & Li, 2020).

Double or multiple scoring by independent raters is a foundational practice in educational

October 27-29, 2025 ©2025 National Council on Measurement in Education (NCME)

measurement for ensuring scoring consistency. However, this approach is costly and time-intensive, requiring the recruitment and training of multiple raters for every item and response (Fliss et al., 1981; Gwet, 2014; Wiggins, 1990).

Alternative cost-saving strategies have emerged to alleviate these resource constraints. One common approach is to double score only a randomly selected subset of responses, though this strategy may be suboptimal when the precise classification of students into performance levels is critical (Finkelman et al., 2009). Alternatively, targeted double scoring (TDS) focuses on responses falling near the critical score range (e.g., pass/fail cutoff), aiming to improve scoring accuracy and reliability (Finkelman et al., 2009; Miao et al., 2023; Sinharay et al., 2022). However, the effectiveness of TDS depends on the accurate identification of the critical score range. Xu and Wind (2025) also found no notable psychometric advantage for TDS over random double-scoring approaches.

Importantly, double or multiple scoring, whether applied to all responses or a subset, substantially increases costs and time compared to single human scoring, creating a persistent tension between scoring quality and practical feasibility. This study explores a novel strategy to optimize reliability scoring within a single human scoring framework, achieving cost-effective and comprehensive measurement without the need for additional human scoring.

3 Method

3.1 Dataset

The PIRLS assesses fourth-grade students' reading comprehension in more than 50 countries globally on a five-year cycle since 2001. In PIRLS 2021, approximately half of the participating countries (*n*=27) transitioned to computer-based testing (digital PIRLS). From the 18 items with reported CCSR values in PIRLS 2021, we selected 2 two-point CR items, using data from all countries participating in digital PIRLS (see Table 1). These two-point items were selected as they are the only two-point "trend" items that will be reused for PIRLS 2026, and this study supports PIRLS 2026 scoring preparation. Notably, one item exhibited the most problematic CCSR of 0.768, making it a challenging yet ideal candidate for validating our new reliability scoring approach.

Item	Process	*N*	CCSR
1	Focus on and retrieve	14,875	0.868
2	Straightforward inferences	14,151	0.768

Table 1: PIRLS trend items used in the study

3.2 Multilingual Response Translation

We utilized a standardized prompt template with GPT-4o to translate non-English responses into English and to rectify spelling and grammatical errors in English responses using GPT-4o (i.e., gpt-4o-2024-08-06). The prompt template incorporated four key components, as detailed in Table 2 (Jung et al., under review). This Zero-Shot-Chain-of-Thought (Zero-Shot-CoT) is task-agnostic (Kojima et al., 2022), enabling its application across diverse items to generate contextually appropriate translations.

Component	Content
Instruction	Comprehensive guidance on AS
Reading passage	A written text serving as the stimulus
Question	A question consisting of one or two sentences
Scoring guide	Rubric for scoring an item, including descriptions and examples

Table 2: PIRLS scoring template components

3.3 Response Flagging and Auto-Scoring

Following translation, we implemented a two-stage data flagging process. First, untranslated responses were flagged as 'missing' and excluded from subsequent analysis. Second, semantically meaningless responses were flagged as 'meaningless', assigned a score of 0, and retained as valid responses for analysis (included in the weighted majority scoring). Detailed criteria for each flagging stage are provided below.

Missing Flagging: Responses were classified as 'missing' if they met either of two criteria: (1) GPT-4o explicitly marked them as 'untranslatable' during translation, or (2) their English vocabulary was less than 75% of tokenized words. This missing flag was only applied to responses exceeding 8 characters. Linguistic preprocessing included lower-casing, lemmatization, and tokenization by spaCy's en_core_web_lg model in Python. The English vocabulary percentage was calculated using the PyEnchant dictionary. Proper nouns (e.g., "California" or "Marie"), identified via spaCy's Named Entity Recognition, counted as valid English vocabulary.

Meaningless Flagging: After excluding missing responses, we flagged 'meaningless' responses if they were: (1) extremely short or (2) semantic outliers. These responses were assigned a score of 0 but retained in the dataset. Very short responses were defined as those with a normalized translation length L_i<0.03, representing the bottom 3% of the length distribution. Translation length was normalized using min-median normalization to mitigate the impact of extreme outliers:

$$L_i = \frac{l_i - \min(l)}{median(l) - \min(l)} \qquad (1)$$

where l_i is the length of the translated response *i*.

Semantic outliers were identified through a multi-faceted assessment. First, responses with a coherence score (C_i) below 0.20 are flagged. C_i was computed as the average cosine similarity between the embedding of response *i* and the embeddings of all other responses, excluding self-similarity:

$$C_i = \frac{1}{N-1} \sum_{i \neq j}^{N} sim(E_i, E_j) \qquad (2)$$

where $sim(E_i, E_j)$ is the cosine similarity between embeddings of response i and *j*. Response embeddings were generated using the Sentence Transformer model (all-MiniLM-L6-v2) in Python.

Second, responses with a meaningfulness score (M_i) below m were also identified as semantic outliers. The meaningfulness threshold m was determined following the hyperparameter optimization. M_i integrates both coherence and normalized length with weights:

$$M_i = 0.80 \times C_i + 0.20 \times L_i \qquad (3)$$

M_i was examined when responses were deemed semantic outliers if the average cosine similarity of their top k most similar responses (as determined during the hyperparameter optimization phase) fell below 0.80.

3.4 Reliability Scoring with Optimal Hyperparameters

Our reliability scoring approach scored responses using a weighted majority scoring algorithm based on cosine similarity between response embeddings.

Similarity Measurement: Response embeddings were generated using the all-MiniLM-L6-v2 model, and cosine similarities were calculated between all response pairs. For each response i, we identified the top k most similar responses based on the highest cosine similarities, where k is a hyperparameter optimized through grid search.

Weighted Majority Scoring: For each response *i*, the majority score s*∈{0, 1, 2} was determined as:

$$s^* = \arg max_s \left(W_{is} = \sum_{j \in S_{is}} sim\left(E_i, E_j\right) \right) \quad (4)$$

where S_{is} is the set of the top *k* similar responses (neighbors) to response *i* with human score *s*. The score $s*$ was assigned only if its proportion of the total weighted score exceeds the weight threshold *WT*, which was optimized via grid search. Otherwise, the response was flagged as 'inconsistent' if the proportion fell below *WT*, indicating that human scores among similar responses varied too widely to assign a reliable majority score.

$$\frac{W_{is^*}}{\sum_s W_{is}} > WT \qquad (5)$$

Hyperparameter Tuning via Grid Search: We conducted a systematic grid search over k ∈ {1, 2, 3, 4, 5, 10, 15} (number of similar responses) and *WT* ∈ {0.60, 0.65, 0.70, 0.75} (weight threshold) to optimize the reliability scoring. All 28 unique hyperparameter combinations were examined using Python's itertools.product.

3.5 Evaluation

The grid search evaluated each hyperparameter combination based on two criteria: (1) minimizing the proportion of responses labeled as 'inconsistent', and (2) maximizing weighted exact agreement (Weighted EA).

Weighted EA quantifies the agreement between human and majority scores, assigning more weight to matches (where human score equals majority score) that exhibit higher cosine similarity. It was calculated as the ratio of the sum of average cosine similarities for responses with matching to the sum of average cosine similarities for all responses. After determining optimal values for *k* and *WT*, several meaningfulness thresholds (*m*) were tested to identify the optimal threshold for detecting semantic outliers. The appropriateness of each threshold was evaluated by analyzing human score distributions, with accurate flagging confirmed by human scores of 0.

Following the hyperparameter optimization, the optimized reliability scoring was analyzed in detail, focusing on the majority score (*s**) distribution and cosine similarity statistics.

4 Results

Hyperparameter Optimization: The grid search results identified the optimal hyperparameter setting as *WT*=0.60 and *k*=3, which minimized the inconsistency proportion and maximized the weighted EA, as detailed in the Appendix. Under this configuration, the inconsistency proportions were very low (0.80% for Item 1 and 2.02% for Item 2), and the weighted EAs (0.881 for Item 1 and 0.755 for Item 2) closely aligned with their corresponding CCSR values (0.868 for Item 1 and 0.768 for Item 2).

Using the optimal hyperparameters (*WT*=0.60 and *k*=3) along with *m* = 0.30, we achieved highly accurate detection of semantic outlier responses, as shown in Tables 3 and 4. For Item 1, 99.40% of responses flagged as 'meaningless' received a human score of 0, compared to 87.08% for Item 2. The reduced detection accuracy for Item 2 was anticipated, as it showed the most significant CCSR issues in PIRLS 2021 (CCSR = 0.768), suggesting inconsistent cross-country scoring, or a higher prevalence of borderline responses susceptible to scoring variations across countries and languages. Given the more reliable performance of Item 1, we adopted *m* = 0.30 for our optimized reliability scoring.

Meaningfulness (*m*)	Human Score (%)		
	0	1	2
0.25	99.02	0.98	0.00
0.26	99.14	0.86	0.00
0.27	99.23	0.77	0.00
0.28	99.30	0.70	0.00
0.29	99.36	0.64	0.00
0.30	99.40	0.60	0.00

Table 3. Human score distribution for 'meaningless' responses to Item 1

Meaningfulness (*m*)	Human Score (%)		
	0	1	2
0.25	93.94	4.94	1.12
0.26	91.86	6.86	1.29
0.27	91.10	7.57	1.33
0.28	90.69	8.10	1.21
0.29	88.13	10.16	1.71
0.30	87.08	10.83	2.09

Table 4. Human score distribution for 'meaningless' responses to Item 2

Reliability Scoring Assessment: First, we examined the majority score distribution (*s**), as presented in Table 5. The average proportions of inconsistent and missing responses were 1.41% (*n*=203) and 1.51% (*n*=218), respectively. This indicates that our reliability scoring approach effectively assigned scores to most responses (97.69% for Item 1 and 96.48% for Item 2) by leveraging their top three most similar neighbors. As expected, Item 2 exhibited a slightly higher inconsistency proportion of 2.02%, consistent with its problematic CCSR. The proportion of missing responses was also low across both items, suggesting that GPT-4o demonstrated a strong capability in translating non-English language responses, including those from low-resource languages such as Arabic, Lithuanian, and Slovak, into English.

Majority score	Item 1		Item 2	
	n	%	*n*	%
0	4314	29.00	5049	35.68
1	3356	22.56	6364	44.97
2	6862	46.13	2240	15.83
Inconsistent	119	0.80	286	2.02
Missing	224	1.51	212	1.50

Table 5. Majority score distribution

Next, we analyzed cosine similarity statistics to assess the effectiveness of our reliability scoring in capturing semantically similar responses, both across all responses and within each response's top three similar neighbors (see Table 6). The mean of average cosine similarities was high, at 0.932 for Item 1 and 0.891 for Item 2, with standard deviations below 0.1, indicating very low variability across responses (see Figures 1 and 2). Additionally, the top three cosine similarities per response tend to be tightly clustered, with very low standard deviation reflecting minimal internal semantic variability among each response's nearest neighbors. These demonstrate the robust performance of our reliability scoring in detecting semantically coherent neighbors.

Item	Mean avg cos sim	SD of avg cos sim	Avg SD of top 3 cos sim
1	0.932	0.098	0.007
2	0.891	0.095	0.012

Table 6. Statistics on average cosine similarity

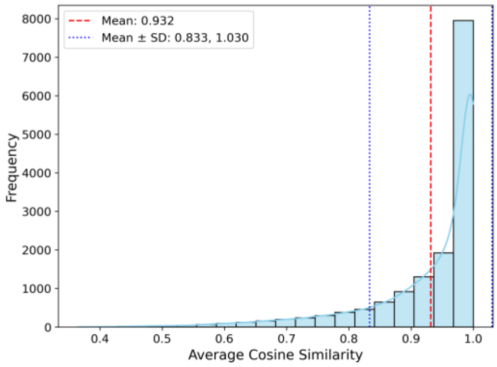

Figure 1. Average cosine similarity for Item 1

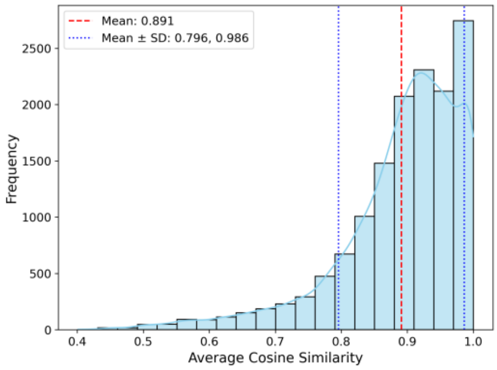

Figure 2. Average cosine similarity for Item 2

5 Discussion

Our findings demonstrate that optimized reliability scoring can effectively evaluate CCSR in multilingual contexts without requiring additional human raters. Although double or multiple scoring has traditionally been the gold standard for achieving consistency (Williamson et al., 2012), prior research (Sinharay et al., 2023; Song & Lee, 2022; Wiggins, 1990) highlights its resource-intensive nature and associated practical and methodological challenges. Our method provides a resource-efficient alternative, utilizing initial human scoring with all responses (over 14,000 responses per item) to achieve results comparable to established CCSR practices. Moreover, this approach enables a comprehensive assessment of individual countries' scoring practices on a global scale using weighted EA or kappa statistics disaggregated by country and language. This facilitates the detection of possible scoring inconsistencies in specific countries or languages and the identification of problematic items (Jung et al., under review).

Despite these promising results, this study has limitations. First, we examined only two two-point "trend" items with available CCSR values, selected for the PIRLS 2026 scoring preparation. Future studies should examine the scalability of this approach across a wider range of item types, including both one- and two-point items. Second, while our approach successfully identified the three most similar neighbors for all responses, responses with low average cosine similarity require further scrutiny. Specifically, responses assigned an initial human score of 2 but exhibiting very low average cosine similarity scores may indicate initial human scoring errors, limitations in our reliability scoring, or both. These cases warrant review by content experts to better understand the sources of scoring discrepancies.

6 Conclusion

This study highlights the effectiveness of optimizing reliability scoring through key hyperparameter optimization and a similarity-aided weighted majority scoring method. This approach robustly measures cross-country consistency by leveraging initial human scoring alongside all responses, offering a more inclusive and cost-effective alternative to existing CCSR. Our novel approach provides a valuable measure for evaluating scoring consistency on a global scale, enabling more accurate and reliable reporting to participating countries.

References

Ercikan, K., & Por, H. H. (2020). Comparability in multilingual and multicultural assessment contexts. Comparability of large-scale educational assessments: Issues and recommendations, 205-225.

Finkelman, M., Darby, M., & Nering, M. (2009). A two-stage scoring method to enhance accuracy of performance level classification. Educational and Psychological Measurement, 69(1), 5-17.

Fleiss, J. L., Levin, B., & Paik, M. C. (1981). The measurement of interrater agreement. Statistical methods for rates and proportions, 2(212-236), 22-23.

Gwet, K. L. (2014). Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters. Advanced Analytics, LLC.

Jung, J. Y., Tyack, L., & von Davier, M. (2025). Towards the Implementation of Automated Scoring in International Large-scale Assessments: Scalability and Quality Control. Computers and Education: Artificial Intelligence, 100375.

Jung, J. Y., Tyack, L., & von Davier, M. (under review). Optimizing Automated Scoring in ILSAs with Prompt Compression Computers and Education: Artificial Intelligence.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. Advances in neural information processing systems, 35, 22199-22213.

Livingston, S. A. (2009). Constructed-Response Test Questions: Why We Use Them; How We Score Them. R&D Connections. Number 11. Educational Testing Service.

Miao, J., Sinharay, S., Kelbaugh, C., Cao, Y., & Wang, W. (2023). Evaluating targeted double scoring for the performance assessment for school leaders using imputation and decision theory. ETS Research Report Series, 2023(1), 1-10.

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. Journal of applied measurement, 4(4), 386-422.

Okubo, T., Houlden, W., Montuoro, P., Reinertsen, N., Tse, C. S., & Bastianic, T. (2023). AI scoring for international large-scale assessments using a deep learning model and multilingual data.

Scully, D. (2017). Constructing multiple-choice items to measure higher-order thinking. Practical Assessment, Research and Evaluation (PARE), 22(1), 1-13.

Sinharay, S., Johnson, M. S., Wang, W., & Miao, J. (2023). Targeted double scoring of performance tasks using a decision-theoretic approach. Applied Psychological Measurement, 47(2), 155-163.

Song, Y. A., & Lee, W. C. (2022). Effects of Using Double Ratings as Item Scores on IRT Proficiency Estimation. Applied Measurement in Education, 35(2), 95-115.

von Davier, M., Mullis, I. V. S., Fishbein, B., & Foy, P. (Eds.). (2023). Methods and Procedures: PIRLS 2021 Technical Report. Boston College, TIMSS & PIRLS International Study Center. https://pirls2021.org/methods

Wang, Y., & Li, S. (2020). Issues, challenges, and future directions for multilingual assessment. Journal of language teaching and research, 11(6), 914-919.

Wiggins, G. (1990). The case for authentic assessment. Practical assessment, research, and evaluation, 2(1).

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated

Xu, Y., & Wind, S. A. (2025). Examining the Psychometric Impact of Targeted and Random Double-Scoring in Mixed-Format Assessments. Educational Measurement: Issues and Practice, 44(1), 18-30.

Yamamoto, K., He, Q., Shin, H. J., & von Davier, M. (2017). Developing a machine-supported coding system for constructed-response items in PISA. ETS Research Report Series, 2017(1), 1-15.

Appendices

A. Grid Search Results

Weight threshold	*k*	Inconsistency (%)	Weighted EA
	1	13.18	0.867
	2	13.14	0.800
0.60	3	0.80	0.881
0.65	3	6.70	0.880
0.70	3	5.27	0.753
0.75	3	5.55	0.753
0.60	4	6.06	0.851
0.65	4	0.97	0.851
0.70	4	6.70	0.851
0.75	4	10.66	0.805
0.60	5	8.70	0.861
0.65	5	8.75	0.825
0.70	5	19.70	0.825
0.75	5	6.70	0.825
0.60	10	10.66	0.861
0.65	10	11.33	0.837
0.70	10	12.76	0.820
0.75	10	19.70	0.788
0.60	15	13.23	0.853
0.65	15	10.66	0.833
0.70	15	15.43	0.806
0.75	15	17.64	0.772

Table 1. Grid search results on Item 1

Weight threshold	*k*	Inconsistency (%)	Weighted EA
	1	25.43	0.738
	2	26.92	0.604
0.60	3	2.02	0.755
0.65	3	2.54	0.753
0.70	3	40.85	0.509
0.75	3	40.88	0.509
0.60	4	18.18	0.670
0.65	4	18.18	0.670
0.70	4	18.19	0.670
0.75	4	30.73	0.584
0.60	5	15.10	0.690
0.65	5	28.49	0.605
0.70	5	28.50	0.605
0.75	5	28.50	0.605
0.60	10	20.01	0.661
0.65	10	28.29	0.604
0.70	10	33.91	0.569
0.75	10	42.51	0.502
0.60	15	22.61	0.643
0.65	15	29.35	0.595
0.70	15	39.31	0.524
0.75	15	48.88	0.451

Table 2. Grid search results on Item 2