Large-scale Assessments in Education

RESEARCH Open Access

Combining machine translation and automated scoring in international large-scale assessments

Ji Yoon Jung[1]*, Lillian Tyack[1] and Matthias von Davier[1]

*Correspondence:
Ji Yoon Jung
jiyoon.jung@bc.edu
[1]Boston College, TIMSS & PIRLS International Study Center, 188 Beacon St., Chestnut Hill, MA 02467, USA

Abstract

Background Artificial intelligence (AI) is rapidly changing communication and technology-driven content creation and is also being used more frequently in education. Despite these advancements, AI-powered automated scoring in international large-scale assessments (ILSAs) remains largely unexplored due to the scoring challenges associated with processing large amounts of multilingual responses. However, due to their low-stakes nature, ILSAs are an ideal ground for innovations and exploring new methodologies.

Methods This study proposes combining state-of-the-art machine translations (i.e., Google Translate & ChatGPT) and artificial neural networks (ANNs) to mitigate two key concerns of human scoring: inconsistency and high expense. We applied AI-based automated scoring to multilingual student responses from eight countries and six different languages, using six constructed response items from TIMSS 2019.

Results Automated scoring displayed comparable performance to human scoring, especially when the ANNs were trained and tested on ChatGPT-translated responses. Furthermore, psychometric characteristics derived from machine scores generally exhibited similarity to those obtained from human scores. These results can be considered as supportive evidence for the validity of automated scoring for survey assessments.

Conclusions This study highlights that automated scoring integrated with the recent machine translation holds great promise for consistent and resource-efficient scoring in ILSAs.

Keywords Automated scoring, Artificial intelligence, Artificial neural networks, Machine translation, Google translate, ChatGPT, International large-scale assessments, TIMSS

Introduction

The Trends in International Mathematics and Science Study (TIMSS) 2019 cycle marked the transition from paper-based to computer-based testing and included more innovative constructed response (CR) items (Martin et al., 2020). In contrast to conventional multiple-choice (MC) items, CR items facilitate deeper and more complete learning by asking students to define a problem, perform investigations, and communicate findings

(Bennett, 1991; Darling-Hammond & Adamson, 2010; Liu et al., 2014). In science education, the use of CR items is encouraged to examine the understanding of core ideas and conduct scientific practices (Zhai et al., 2020). However, the wider use of CR items in international large-scale assessments (ILSAs) has been limited due to the resource-intensive nature of human scoring and the challenges for reliable and accurate scoring of huge volumes of multilingual student responses (Yamamoto et al., 2017).

The human scoring of CR items is known to be expensive, time-consuming, and labor-intensive. Bennett (1991) stated that the operational cost and efforts associated with CR items are generally more substantial than traditional MC items. This disparity has become even wider with advances in computer-based data collection and machine scoring of selected response (for example MC) items using statistical programming languages such as SAS and R. In addition, the recruitment of professional human raters, rigorous training, and continuous monitoring are needed to achieve a high level of consistency and accuracy (Ramineni & Williamson, 2013; Zhang, 2013). Even with intensive training and monitoring, scoring issues derived from fatigue, distraction, and rater effects like severity and leniency still occur (McClellan, 2010; Myford & Wolfe, 2009; von Davier et al., 2023; Wolfe & McVay, 2012).

Although the TIMSS & PIRLS International Study Center provides detailed explanations of scoring rubrics and extensive training to mitigate such risks, achieving a high level of scoring reliability involves a significant workload and expense for participating countries. Among the many challenges, human raters in participating countries must be trained with scoring materials translated into their native language(s) by head-scorers or scoring trainers who attended an international scoring training where materials were provided in English (Martin et al., 2020). Therefore, scoring large volumes of multilingual responses is subject to potential scoring inconsistencies not only across countries but also across raters within each country.

The current study follows a common strategy in multilingual natural language processing (NLP), employing machine translation (MT) to translate various non-English languages into English (Balahur & Turchi, 2012; Lucas et al., 2015; Montalvo et al., 2015). Automated scoring in ILSAs is thought to be more challenging than in the monolingual contexts since most NLP tools and research are predominantly focused on English (Hovy & Prabhumoye 2021). In the past few years, MT has advanced significantly. META's artificial intelligence (AI) model, No Language Left Behind, produces high-quality translations for 200 different languages (META, 2022). Google Translate supports 133 languages, including 24 low-resource languages (Caswell, 2022). OpenAI's Generative Pre-trained Transformer (GPT) models also emerged as excellent translators, generating contextually relevant translations (Hendy et al., 2023; Timothy, 2023). Jiao et al. (2023) found that ChatGPT competes well with commercial translation engines, especially for high-resource languages.

In addition to MT, this study proposes using the Bag-of-Words (BoW) to score CR items requiring very short answers in ILSAs. The BoW identifies unique words (features) within the data and counts the frequency of each word in individual texts. Although the BoW representation is often criticized for its sparsity, high dimensionality, and challenges in capturing complex meanings, it can be a suitable approach for scoring CR items that ask for brief answers including key concepts. In the TIMSS items selected for this study, students often provide succinct answers with fourth-grade level words and their responses have many identical keywords, which is one of the features of simple CR items (Yamamoto et al., 2017). This characteristic contributes to the lower sparsity and dimensionality of the BoW representation, suggesting that BoW can efficiently extract crucial keywords to classify correct and incorrect responses. de Vries et al. (2018) advocate for the utility of combining BoW with MT for text analysis in a multilingual context. Also, the verifiable key features of BoW enable subject-domain experts to review whether the features used for automated scoring align with the established rubric. More importantly, using the common key features in all responses helps mitigate possible scoring inconsistencies across countries and languages.

Despite the considerable interest in automated scoring, most studies have focused on applications in the monolingual context. This study aims to show that the combination of automated scoring and MT can be a useful support for or even an alternative to human scoring in ILSAs involving diverse countries and languages. This study addresses the following questions:

1. Can automated scoring achieve comparable performance to human scoring across different countries and languages without compromising the psychometric properties of items?
2. Does MT appropriately convert non-English language responses into English to construct a unified cross-lingual automated scoring model?
3. What are the sources of misalignment between human and automated scoring?

Background

There has been a long desire to apply automated scoring in education. Starting with Ellis Page's first automated scoring engine (Page, 1966), early research dates back to the late 1960s. Recent advances in digital data collection, NLP, machine learning algorithms, computer software, and hardware have enabled the operational use of automated scoring in multiple assessment programs (Foltz et al., 2020) such as ETS's e-rater, Duolingo's English Test, and Pearson's Intelligent Essay Assessor. Despite these accomplishments, the use of automated scoring in multilingual contexts is still lacking. The fundamental difficulty in multilingual automated scoring is to ensure consistent and accurate scoring of a vast number of responses across all the languages in which ILSAs are administered. Given that the 2019 cycle of TIMSS collected data from 64 countries written in 50 different languages (Martin et al., 2020), the application of automated scoring in ILSAs may be considered challenging.

While the initial concept of MT was proposed by Warren Weaver in 1947, MT has shown significant improvement with the advent of neural networks (Britz et al., 2017; Hutchins, 2007; Wang et al., 2021). Recent MT engines provide fast, accurate, and affordable translation with minimal or no loss of meaning. To tackle multilingual responses in ILSAs, we chose to use MT and construct a unified model for all languages instead of developing separate models for each language. This cross-lingual model alleviates the laborious task of collecting and building training sets for individual languages, especially those with low resources. Although monitoring translation quality is crucial, achieving a 'perfect' translation is not the primary goal. Rather, our focus is on demonstrating that machine-translated responses can be automatically scored with an accuracy level equivalent to or surpassing that of non-translated responses (i.e., English language responses).

Moreover, the abundance of responses collected in ILSAs has historically posed a challenge for scoring CR items. Modern NLP and artificial neural networks (ANNs) can easily handle large datasets due to powerful computer algorithms. Unlike early machine scoring from the mid-to-late 1900s, which was impractical for ILSAs due to their reliance on manual feature selection and rule-based techniques (Cahill & Evanini, 2020), contemporary AI models can automatically learn patterns and rules from the data, saving both time and labor. ANNs are more extensive and flexible compared to previous machine-supported scoring and can be applied to various tasks, including automated scoring, text classification, paraphrasing, language generation, and question-answering (Abiodun et al., 2018; Kim, 2014; Mallinson et al., 2017; Prakash et al., 2016; Prasanna & Rao, 2018; Sutskever et al., 2011; Wang & Jiang, 2016).

This study aims to investigate the performance of AI-powered automated scoring in ILSAs, with a focus on the application of MT in scoring short CR items.

Methods

Data

The current study used six short CR items from TIMSS 2019. These items are homogenous in terms of the subject domain (science), target students (fourth-grade students), dichotomous scoring (correct response = 1; incorrect response = 0), and the elicitation of very short responses. We analyzed the multilingual student responses involving eight countries and six different languages: four Latin alphabet languages (German, French, Turkish, and English) and two non-Latin alphabet languages (Chinese and Korean). These countries and languages were selected to examine whether automated scoring could perform consistently across different types of languages. The selection of languages was also based on the availability of native speakers of these languages working at the TIMSS & PIRLS International Study Center, where this study was conducted. The item-by-country sample sizes are shown in Table 1. Detailed data information can be shared upon request.

The student responses were very succinct—after translation, they averaged 33 characters with Google Translate and 36 characters with ChatGPT. This is notably short in comparison to the common definition of short texts, which have a maximum length of 200 characters (Song et al., 2014). The range of response lengths varied from 18 to 57 characters for Google Translate and 23 to 57 characters for ChatGPT. Interestingly, C5, an English-speaking country, had the lengthiest average responses, ranging from 43 to 61 characters across all six items.

Table 1 Item-by-country sample size

Item	C1	C2	C3	C4	C5	C6	C7	C8	Total
Item 1	535	406	462	540	1,204	351	489	530	4,517
Item 2	492	392	459	532	1,208	361	481	538	4,463
Item 3	593	459	528	565	1,208	360	518	549	4,780
Item 4	562	437	482	544	1,194	337	488	539	4,583
Item 5	543	434	461	545	1,214	368	518	536	4,619
Item 6	625	447	531	547	1,205	373	536	551	4,815

Note C1 & C2 = German-speaking countries; C3 = French-speaking country; C4 = Turkish-speaking country; C5 = English-speaking country; C6 & C7 = Chinese-speaking countries; C8 = Korean-speaking country

Procedures

Data partitioning

The data was split into training and test sets at a ratio of 80:20. Within the training set (80% of the whole data), cross-validation (CV) was performed, using 80% for training and 20% for validating the model's performance. The test set (20% of the whole data) is independent and previously unseen data. During the data split, we assigned a subset of double-scored responses to the training set. This subset of responses was derived from 200 randomly selected responses per country, which were scored by two independent human raters during TIMSS 2019 data collection. We duplicated responses that received consistent scores from both human raters while excluding responses with conflicting scores. This approach aimed to include more reliable responses into the training set and thus construct more accurate ANNs (Ilse et al., 2018). Sample sizes for the multilingual training set and individual countries' test set are shown in Table 2.

Multiple MT

We employed Google Translate API and ChatGPT API (i.e., gpt-3.5-turbo) to translate non-English language responses into English using Python (version 3.11.4). Google Translate is a translation engine supporting more than 100 language pairs that uses a pre-trained neural MT model (Google, 2023a). It automatically detects the source language and translates non-English responses into English. ChatGPT, on the other hand, is a large language model that uses self-attention mechanisms to produce context-based natural language responses. It is the most powerful and cost-effective model in the GPT-3.5 models (OpenAI, 2023a). We instructed ChatGPT to translate a given non-English language response into English considering the context (i.e., the English stem/question of the item). Incorporating the context in the prompt directed ChatGPT to generate a more question-relevant translation rather than a translation without context, which could take a different off-topic direction if responses were unclear, short, or both.

MT enabled the ANNs to be trained and tested on very large English-only data that includes both native English responses and non-English responses translated into English. The advantage of multiple MT is that it can lead to improved translation quality rather than relying on a single translation engine. We aimed to select a more suitable MT tool between Google Translate and ChatGPT for more accurate automated scoring. The evaluation of MT quality is goal-oriented. The aim is to select MT that extracts useful features applicable to all languages, rather than solely focusing on perfect translation. Obtaining common BoW features is possible when the key concepts in a variety of languages are appropriately transformed by the translation engine of choice.

Table 2 Sample size for multilingual training set (80%) and individual country's test set (20%)

Item	Training	Test								
		C1	C2	C3	C4	C5	C6	C7	C8	Total
Item 1	4,216	144	117	129	147	279	106	133	143	1,198
Item 2	4,236	137	118	130	144	279	108	135	147	1,198
Item 3	4,411	158	130	145	153	278	106	142	149	1,261
Item 4	4,232	150	127	131	147	275	102	132	144	1,208
Item 5	4,328	146	126	129	148	280	109	143	146	1,227
Item 6	4,412	163	129	145	148	277	109	145	150	1,266

Note C1 & C2 = German-speaking countries; C3 = French-speaking country; C4 = Turkish-speaking country; C5 = English-speaking country; C6 & C7 = Chinese-speaking countries; C8 = Korean-speaking country

Pre-processing, BoW, & ANNs

We applied the identical pre-processing, BoW, and ANN procedures to two sets of translated responses: (1) Google-translated responses with native English responses, and (2) ChatGPT-translated responses with native English responses. The preferred MT was determined using the test set, considering average human-machine score agreements and the log odds ratio for individual countries (described further in the results).

Common NLP pre-processing steps were applied such as tokenization, lower-casing, spelling correction, and stemming (reducing a word to its stem). NLP tools such as *NLTK* and *pyspellchecker* in Python were used. Regarding spelling correction, we replaced misspelled words in the test set with correct words elicited from the training set (Jung et al., 2022). For English-speaking countries (here only C5), an additional spelling correction was implemented by replacing misspelled words with the first suggested word from *pyspellchecker.* The rationale for this additional step was that many non-English misspelled words were corrected during MT, while English responses, which did not undergo MT, were left with more misspelled words. Following spelling correction, we only maintained words appearing at least 0.05% in the training set to exclude any irrelevant words in the feature matrix. The BoW represented all translated responses and English language responses within a common key feature matrix. For example, the BoW can transform a student response, "*because weather is cold*", into {"*because*", "*weather*", "*is*", "*cold*"}, which could be projected to {0, 1, 1, 1}.

Next, Fully-connected feed-forward neural networks (FNNs) were implemented using the *sklearn* package in Python. Being structured into the input, hidden, and output layers, FNNs have no cyclic connections between layers, and all the neurons in successive layers are connected. They are frequently used in practical applications because of their fast learning speed and acceptable performance (Han et al., 2019; Le & Huynh, 2016). The BoW key features were fed to the input layer and then processed through the hidden layer and output layers. Machine scores of 1 and 0 were represented in the output layer. We performed a 5-fold CV on the training set to select the most optimized values of hyperparameters, such as the number of hidden neurons. CV is a widely used technique in machine learning to assess the capability of models to generalize their predictions to new data and prevent overfitting (Berrar, 2019). We trained the FNNs on 80% of the training set and tested them on 20% of the training set (validation or development set). The final FNN was then applied to the unseen test set.

Evaluation Metrics

The evaluation of automated scoring performance included standard text classification metrics such as the exact match ratio, Cohen's kappa (κ), F1 scores, and standardized mean score difference (SMD). Additionally, translation performance between Google Translate and ChatGPT was evaluated using the log odds ratio (LOR). Psychometric measures, including adjusted item-total correlations (AITC) and item difficulty, were used to examine the impact of automated scoring on the psychometric quality of the items.

Exact match ratio. The exact match ratio, a widely used metric, quantifies the proportion of agreement between machine and human scores. Instances where human and machine scores perfectly aligned were categorized as Both Incorrect (BI) and Both

Correct (BC) response pairs. Disagreements were represented by Disagrees (D1 and D2) in pairs (see Table 3).

$$Exact\ Match\ Ratio \quad = \frac{BI + BC}{BI + BC + D1 + D2}$$

Cohen's Kappa. Cohen's kappa is considered a more robust measure than the exact match ratio, as it evaluates inter-rater agreement beyond chance. A kappa of 0 indicates an agreement equivalent to chance. We opted for the standards set by Landis and Koch (1977): values ≤ 0.00 classified as poor, 0.00-0.20 as slight, 0.21–0.40 as fair, 0.41–0.60 as moderate, 0.61–0.80 as good, and 0.81-1.00 as very good agreement. Liu et al. (2014) also applied these criteria to assess their automated scoring engine, which was used for scoring low-stake items. We assessed inter-rater reliability using the kappa statistic for both human-human and human-machine scoring.

$$Kappa = \frac{P_{observation} - P_{chance}}{1 - P_{chance}}$$

F1 scores. F1 scores serve as a crucial metric, especially for imbalanced data, as they measure the harmonic mean of precision and recall, ranging from 0 to 1. Higher F1 scores indicate low false negatives (D1) and false positives (D2), implying lower human-machine score disagreements in this study (refer to Table 3). Given the uneven class distribution (correct vs. incorrect) in some items in our study, F1 scores provide a more accurate representation of automated scoring performance.

$$Precision = \frac{BC}{BC + D2}$$

$$Recall = \frac{BC}{BC + D1}$$

$$F1\, score = \frac{2 \times BC}{2 \times BC + D1 + D2} = \frac{2 \times (precision \times recall)}{(precision + recall)}$$

SMD. SMD refers to the mean score difference between human and machine scores divided by the pooled standard deviations. An SMD of 0 indicates that there is no difference between human and machine scores. Positive values mean that automated scoring yields a higher mean score than human scoring while negative values indicate the opposite. SMD is also a good metric to assess the discrepancy in item difficulty between human and machine scores. Williamson et al. (2012) suggested using a threshold of 0.15 to indicate a satisfactory level of agreement. SMD is calculated as below:

$$SMD = \frac{\overline{X}_M - \overline{X}_H}{\sqrt{(S_M^2 + S_H^2)/2}}$$

Table 3 Confusion matrix in automated scoring

		Human Score	
		Incorrect (0)	Correct (1)
Machine Score	Incorrect (0)	Both Incorrect (BI)	Disagree (D1)
	Correct (1)	Disagree (D2)	Both Correct (BC)

where $\overline{X}_M$ and $\overline{X}_H$ are the mean of machine and human scores, respectively. S_M^2 and S_H^2 are the variance of machine and human scores, respectively.

LOR. The odds ratio compares two sets of odds, representing the ratios of the probability of an event occurring to the probability of it not occurring. In this study, an event occurring signifies a match between the machine score and the human score. We calculated the odds for the exact match ratio in both Google and ChatGPT-translated data using a logarithmic scale. A LOR value of 0 means that the exact match ratio derived from Google and ChatGPT-translated data is the same, indicating an equivalent translation effect. A negative LOR, resulting from a greater exact match ratio in Google-translated data compared to ChatGPT data, implies that Google Translate provides more appropriate translations for automated scoring. Conversely, a positive LOR implies that ChatGPT provides more suitable translations than Google Translate.

$$LOR = LN\left(\frac{P_{ChatGPT}/(1 - P_{ChatGPT})}{P_{Google}/(1 - P_{Google})}\right)$$

AITC. AITC is the correlation between each item and the total score, excluding the item of interest. This correlation was employed to prevent biased estimation. In TIMSS 2019, items are grouped into 14 blocks consisting of 10 to 14 items (Mullis & Martin, 2017). In each scoring method (i.e., human and automated scoring), the AITC was calculated by assessing the correlation between each item and the percentage of correct responses within the item's block, excluding the item itself.

Item difficulty. Item difficulty measures the percentage of correct responses, with lower values indicating more challenging items. We computed item-by-country difficulty using both human and machine scores to explore whether different scoring methods influenced item difficulty.

Results

Reliability of human scoring

Human-human inter-rater reliability was computed using the double-scored responses from the within-country reliability scoring sample. Human raters showed high to perfect agreements, with kappa values ranging from 0.84 to 1.00 across items and countries. These values indicate the high reliability of human scoring. Notably, C6 consistently reached perfect inter-rater reliability for all items. This perfect inter-rater reliability was consistently observed in all other CR items for fourth graders in TIMSS 2019. This might be attributed to a potential misunderstanding of double-scoring, wherein human raters are not permitted to discuss discrepancies to establish a consensus See Table 4.

Table 4 Item-by-country kappa (human-human inter-rater reliability)

Item	C1	C2	C3	C4	C5	C6	C7	C8	Average
Item 1	0.94	0.89	0.93	0.99	0.97	1.00	0.89	0.93	0.94
Item 2	0.98	0.98	0.95	0.94	0.94	1.00	0.98	0.99	0.97
Item 3	0.97	0.94	0.98	1.00	0.90	1.00	0.98	0.99	0.97
Item 4	0.95	0.99	0.84	0.97	0.89	1.00	0.85	0.86	0.92
Item 5	0.88	0.98	0.91	0.98	0.94	1.00	1.00	0.94	0.95
Item 6	0.97	0.98	0.96	0.99	0.91	1.00	0.96	1.00	0.97
Average	0.95	0.96	0.93	0.98	0.93	1.00	0.94	0.95	0.95

Note C1 & C2 = German-speaking countries; C3 = French-speaking country; C4 = Turkish-speaking country; C5 = English-speaking country; C6 & C7 = Chinese-speaking countries; C8 = Korean-speaking country

Table 5 Item-by-country exact match ratio (google translate)

	C1	C2	C3	C4	C5	C6	C7	C8	Average
Item 1	0.90	0.90	0.91	0.89	0.89	0.82	0.89	0.85	0.88
Item 2	0.93	0.97	0.95	0.95	0.95	0.98	0.93	0.96	0.95
Item 3	0.97	0.95	0.96	0.95	0.93	0.98	0.97	0.97	0.96
Item 4	0.92	0.94	0.89	0.90	0.89	0.79	0.89	0.88	0.89
Item 5	0.92	0.94	0.82	0.84	0.85	0.86	0.92	0.90	0.88
Item 6	0.96	0.96	0.88	0.88	0.96	0.83	0.96	0.91	0.92
Average	0.93	0.94	0.90	0.90	0.91	0.88	0.93	0.91	0.91

Note C1 & C2=German-speaking countries; C3=French-speaking country; C4=Turkish-speaking country; C5=English-speaking country; C6 & C7=Chinese-speaking countries; C8=Korean-speaking country

Table 6 Item-by-country exact match ratio (ChatGPT)

	C1	C2	C3	C4	C5	C6	C7	C8	Average
Item 1	0.88	0.91	0.94	0.90	0.92	0.92	0.93	0.87	0.91
Item 2	0.92	0.99	0.95	0.95	0.95	0.97	0.94	0.95	0.95
Item 3	0.98	0.98	0.97	0.95	0.92	0.99	0.97	0.97	0.97
Item 4	0.95	0.93	0.85	0.90	0.88	0.86	0.94	0.90	0.90
Item 5	0.92	0.90	0.85	0.86	0.85	0.92	0.92	0.92	0.89
Item 6	0.98	0.96	0.94	0.91	0.96	0.77	0.94	0.87	0.92
Average	0.94	0.95	0.92	0.91	0.91	0.91	0.94	0.91	0.92

Note C1 & C2=German-speaking countries; C3=French-speaking country; C4=Turkish-speaking country; C5=English-speaking country; C6 & C7=Chinese-speaking countries; C8=Korean-speaking country

Table 7 Item-by-country LOR

Item	C1	C2	C3	C4	C5	C6	C7	C8	Average
Item 1	-0.20	0.12	0.44	0.11	0.35	0.93	0.50	0.17	0.32
Item 2	-0.14	1.12	0.00	0.00	0.00	-0.42	0.16	-0.23	0.00
Item 3	0.42	0.95	0.30	0.00	-0.14	0.70	0.00	0.00	0.30
Item 4	0.50	-0.16	-0.36	0.00	-0.10	0.49	0.66	0.20	0.11
Item 5	0.00	-0.55	0.22	0.16	0.00	0.63	0.00	0.25	0.10
Item 6	0.71	0.00	0.76	0.32	0.00	-0.38	-0.43	-0.41	0.00
Average	0.16	0.19	0.25	0.12	0.00	0.32	0.16	0.00	0.13

Note 1 LOR=Log odds ratio

Note 2 C1 & C2=German-speaking countries; C3=French-speaking country; C4=Turkish-speaking country; C5=English-speaking country; C6 & C7=Chinese-speaking countries; C8=Korean-speaking country

MT selection for automated scoring

High human-machine score agreements were observed across the items and countries for both Google Translate and ChatGPT (see Tables 5 and 6), although automated scoring exhibited slightly superior performance on ChatGPT-translated data. The ChatGPT MT method consistently achieved agreements exceeding 0.85, except for C6 in item 6 (0.77). The lower agreement for this country and item is further explored in the discussion.

Next, the performance of Google Translate and ChatGPT API was assessed using LOR. Although the overall translation quality appears comparable, ChatGPT demonstrates superior performance, as indicated by more positive LORs (see Table 7). ChatGPT was particularly useful for misspelled responses where context (question) plays a crucial role in the translation. Hence, we opted for ChatGPT as our preferred MT tool and proceeded to evaluate the performance of automated scoring based on ChatGPT-translated data.

Table 8 Item-by-country F1 scores

	C1	C2	C3	C4	C5	C6	C7	C8	Average
Item 1	0.81	0.88	0.93	0.88	0.88	0.92	0.87	0.86	0.88
Item 2	0.90	0.99	0.95	0.96	0.94	0.97	0.93	0.94	0.95
Item 3	0.99	0.99	0.98	0.97	0.95	0.99	0.97	0.95	0.97
Item 4	0.94	0.91	0.90	0.87	0.91	0.92	0.96	0.75	0.90
Item 5	0.67	0.71	0.77	0.68	0.68	0.74	0.40	0.74	0.67
Item 6	0.97	0.92	0.95	0.93	0.94	0.81	0.90	0.74	0.90
Average	0.88	0.90	0.91	0.88	0.88	0.89	0.84	0.83	0.88

Note C1 & C2 = German-speaking countries; C3 = French-speaking country; C4 = Turkish-speaking country; C5 = English-speaking country; C6 & C7 = Chinese-speaking countries; C8 = Korean-speaking country

Table 9 Item-by-country kappa (human-machine inter-rater reliability)

	C1	C2	C3	C4	C5	C6	C7	C8	Average
Item 1	0.72	0.81	0.87	0.80	0.82	0.83	0.82	0.73	0.80
Item 2	0.83	0.98	0.91	0.89	0.90	0.94	0.88	0.90	0.90
Item 3	0.94	0.93	0.92	0.87	0.80	0.97	0.94	0.93	0.91
Item 4	0.89	0.85	0.60	0.78	0.71	0.53	0.84	0.69	0.74
Item 5	0.62	0.65	0.66	0.59	0.58	0.70	0.36	0.70	0.61
Item 6	0.95	0.90	0.89	0.81	0.91	0.53	0.86	0.65	0.81
Average	0.82	0.85	0.81	0.79	0.79	0.75	0.78	0.77	0.80

Note C1 & C2 = German-speaking countries; C3 = French-speaking country; C4 = Turkish-speaking country; C5 = English-speaking country; C6 & C7 = Chinese-speaking countries; C8 = Korean-speaking country

Table 10 Item-by-country SMD

	C1	C2	C3	C4	C5	C6	C7	C8	Average
Item 1	0.06	-0.04	0.03	-0.14	0.02	-0.06	0.05	0.04	-0.01
Item 2	0.04	-0.02	-0.03	-0.04	-0.05	-0.02	0.03	0.01	-0.01
Item 3	-0.02	-0.02	0.03	-0.08	-0.09	-0.03	0.00	-0.03	-0.03
Item 4	-0.08	-0.11	-0.10	-0.07	-0.16	0.15	-0.03	-0.05	-0.06
Item 5	0.02	-0.10	0.05	-0.25	-0.06	-0.22	0.11	0.02	-0.05
Item 6	0.02	-0.02	-0.05	-0.07	-0.01	-0.32	-0.09	0.03	-0.06
Average	0.01	-0.05	-0.01	-0.11	-0.06	-0.08	0.01	0.00	-0.04

Note 1 SMD = Standardized mean score difference

Note 2 C1 & C2 = German-speaking countries; C3 = French-speaking country; C4 = Turkish-speaking country; C5 = English-speaking country; C6 & C7 = Chinese-speaking countries; C8 = Korean-speaking country

Comparability of automated scoring to human scoring

Automated scoring using MT demonstrated comparable performance to human scoring across multiple metrics. Machine scores demonstrated good agreement with human scores, with average F1 score and kappa of 0.88 and 0.80, respectively (see Tables 8 and 9). Machine scoring was slightly stricter than human scoring with an average SMD of -0.04 but the difference was marginal (see Table 10). However, Item 5 in C1, C4, C5, and C7 exhibited relatively moderate-to-low values for both F1 scores, ranging from 0.40 to 0.68, and kappa, ranging from 0.36 to 0.62. Item 6 in C6 also displayed a relatively low kappa of 0.53 and a substantial SMD of -0.32, a pattern also flagged in the moderate exact match ratio of 0.77. Performance on items 5 and 6 will be explored further in the discussion.

Impact of automated scoring on psychometric properties

Both human and machine scoring demonstrated good AITC across items on average, with a slightly higher value for human scoring (r_{human} = 0.35; $r_{machine}$= 0.33) (see Table 11).

The AITC generally displayed consistent patterns across countries and scoring methods, with a slightly stronger correlation in human scoring (see Table 12). Particularly noteworthy is Item 6 in C6, where the item-total correlations consistently remained high in automated scoring (r_{human} = 0.44; $r_{machine}$= 0.44), despite being flagged by other metrics such as the moderate exact match ratio value (0.77), moderate kappa value (0.53), and large SMD (-0.32). These results suggest that while automated scoring may be stricter than or deviate from human scoring, the common gold standard, it does not necessarily compromise the item's contribution to the instrument or internal consistency. Such discrepancies do not necessarily indicate errors in automated scoring but could point to potential errors or challenges within the human scoring process. This will be further discussed in the discussion.

Moreover, we observed that AITC can be different within the same language countries depending on the scoring method. This pattern was notable for Item 5 in German-speaking countries (C1 and C2) and Chinese-speaking countries (C6 and C7). In C1, human scores showed higher AITC (r_{human}= 0.23), while in C2, machine scores displayed higher AITC ($r_{machine}$= 0.23). Similarly, in C6, the AITC was higher with human scores (r_{human}= 0.27) whereas in C7, the reverse was true ($r_{machine}$= 0.32). Also, machine scores can even yield higher AITC within the same language countries. For Item 1, the AITC was similar between human and machine scores in C6 (0.30), but the AITC was noticeably higher with machine scores in C7 ($r_{machine}$= 0.23 > r_{human}= 0.13).

Next, the overall patterns of country-by-item difficulty remained consistent across the scoring methods (see Figs. 1, 2, 3, 4, 5 and 6). Importantly, even uncommon patterns were maintained across the scoring method (refer to Fig. 3). In human scoring, Item 3 was relatively easy for students in C7 (r_{human} = 0.54) and C8 (r_{human} = 0.72), while challenging for the other countries, as indicated by item difficulties below 0.30. This distinctive pattern was also similarly reflected in the automated scoring: C7 ($r_{machine}$ = 0.54) and C8 ($r_{machine}$ = 0.71) showed a high percentage of correct responses, whereas the other countries reported low values below 0.25. Yet, we observed noticeable gaps between human and machine scores for C2, C5, and C6 in Item 4, and for C4 and C6 in Item 5. Particularly, C6 consistently showed a gap of 0.06, 0.08, and 0.16 for Items 4, 5, and 6, respectively. These disparities will be further examined in the discussion.

Table 11 Item-by-scoring method AITC

	Human Score	Machine Score
Item 1	0.38	0.36
Item 2	0.33	0.32
Item 3	0.36	0.36
Item 4	0.34	0.31
Item 5	0.26	0.22
Item 6	0.45	0.41
Average	0.35	0.33

Note AITC = Adjusted Item-total Correlation

Table 12 Item-by-country AITC

	C1		C2		C3		C4		C5		C6		C7		C8	
	H	M	H	M	H	M	H	M	H	M	H	M	H	M	H	M
Item 1	0.45	0.45	0.35	0.34	0.51	0.46	0.38	0.41	0.49	0.39	0.30	0.30	0.13	0.23	0.32	0.30
Item 2	0.24	0.32	0.36	0.36	0.40	0.36	0.24	0.22	0.49	0.49	0.32	0.28	0.30	0.30	0.32	0.24
Item 3	0.22	0.22	0.24	0.25	0.25	0.20	0.43	0.42	0.32	0.31	0.14	0.13	0.47	0.43	0.44	0.38
Item 4	0.32	0.31	0.39	0.25	0.32	0.35	0.46	0.48	0.40	0.31	0.28	0.27	0.20	0.16	0.22	0.19
Item 5	0.23	0.15	0.16	0.23	0.28	0.23	0.38	0.19	0.31	0.23	0.27	0.25	0.14	0.32	0.22	0.24
Item 6	0.50	0.48	0.43	0.42	0.44	0.39	0.45	0.40	0.49	0.45	0.44	0.44	0.40	0.42	0.32	0.17
Average	0.33	0.32	0.32	0.31	0.37	0.33	0.39	0.35	0.42	0.36	0.29	0.28	0.27	0.31	0.31	0.26

Note 1 AITC = Adjusted item total correlation

Note 2 H = Human score; M = Machine score

Note 3 C1 & C2 = German-speaking countries; C3 = French-speaking country; C4 = Turkish-speaking country; C5 = English-speaking country; C6 & C7 = Chinese-speaking countries; C8 = Korean-speaking country

Discussion

Toward consistent and resource-efficient scoring

The present study found that automated scoring has great potential for supporting and even possibly replacing the need for labor-intensive human scoring in multilingual contexts. Despite the small test sample size, the automated scoring resulted in generally good agreements between human and machine scores without negatively affecting psychometric characteristics. This finding implies that MT effectively extracted common BoW key features that could be used in all countries and languages while retaining the core meaning. While human scores can vary depending on human rater understanding and biases, automated scoring using shared key features could help reduce scoring inconsistencies within or between countries.

Moreover, automated scoring can significantly reduce the expenses associated with human scoring. Human scoring of multilingual responses in ILSAs necessitates substantial costs, time, and labor. In contrast, the application of automated scoring was remarkably cost-effective and time-efficient. Regarding MT, Google Translate costs $20 per one million characters (Google, 2023b), and ChatGPT $0.002 per 1,000 tokens (around 750 words) (OpenAI, 2023b). MT per student response took 0.14 and 0.42 s by Google Translate and ChatGPT, respectively. Running the ANNs per item took approximately 7.50 min via Python. Considering that inconsistency and high expenses are the fundamental challenges of human scoring, this study suggests that automated scoring may soon be an efficient alternative to human scoring, and allow for more reliable and consistent scoring of CR items in ILSAs.

Misalignment between human and automated scoring

To better understand the nature of misclassified responses, we investigated the likely sources of the human-machine score disagreement. The potential causes we considered are three-fold: (1) errors in automated scoring, (2) errors in human scoring, and (3) true score uncertainty.

First, errors in automated scoring refer to instances where the machine classified responses as incorrect (machine score 0), whereas a human rater classified them as correct (human score 1) (refer to D1 in Table 3). Regarding the BoW approach, we found the lexical diversity of correct responses is one important source of error. Although most correct student responses are homogeneous in this study, we found that the correct answer to Item 4 can be expressed in multiple ways. For instance, the keyword of Item 4 was *sieve* – which was found to be expressed by students as *a bucket with holes*, *colander*, *drainer*, *filter*, *net*, *strainer*, *sifter*, *separator*, *wire mesh*, etc. The BoW did not capture these low-frequency keywords in its feature matrix, but human raters accurately scored a variety of responses as long as they conveyed similar concepts. In future studies, advanced NLP models such as word embedding (e.g., the WordNet-based lemmatization) could be used to identify and address a variety of synonyms (Chen et al., 2019; Mikolov et al., 2013).

Next, errors in human scoring indicate instances where a human rater classified responses as incorrect (human score 0), whereas the machine classified them as correct (machine score 1) (see D2 in Table 3). Humans are not perfect, and therefore, human scores could be inconsistent or inaccurate. Although the inter-rater reliability of human scoring was very high ($\kappa = 0.97–1.00$) in this study, we observed slight within-country

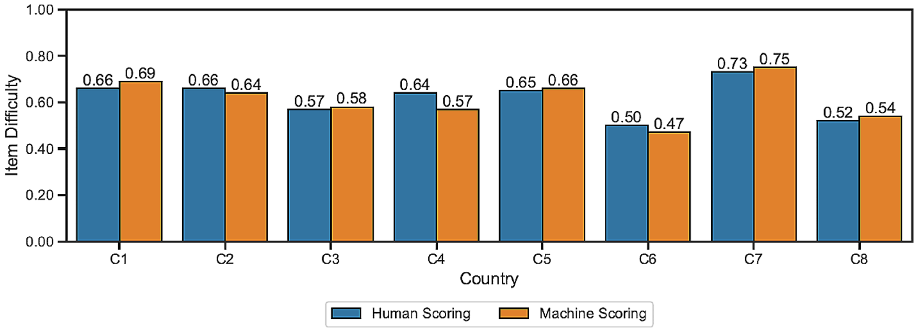

Fig. 1 County-by-item difficulty of item 1. *Note* C1 & C2 = German-speaking countries; C3 = French-speaking country; C4 = Turkish-speaking country; C5 = English-speaking country; C6 & C7 = Chinese-speaking countries; C8 = Korean-speaking country

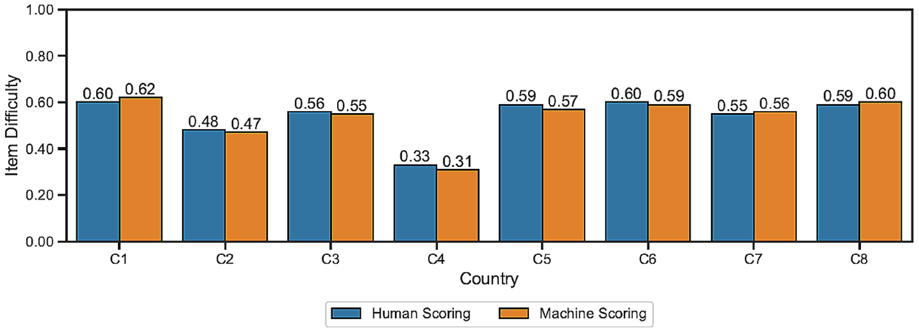

Fig. 2 Country-by-item difficulty of item 2. *Note* C1 & C2 = German-speaking countries; C3 = French-speaking country; C4 = Turkish-speaking country; C5 = English-speaking country; C6 & C7 = Chinese-speaking countries; C8 = Korean-speaking country

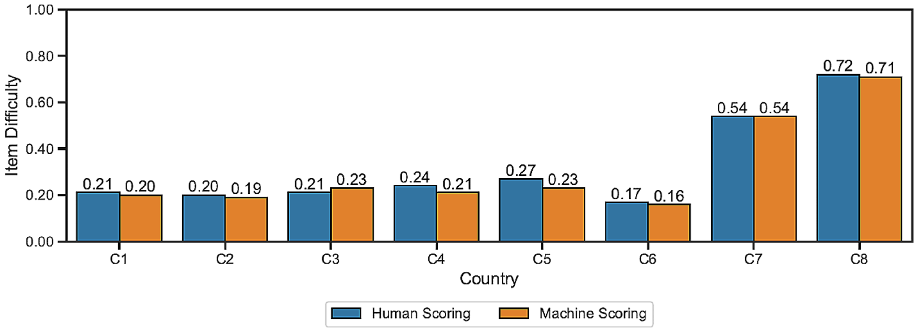

Fig. 3 Country-by-item difficulty of item 3. *Note* C1 & C2 = German-speaking countries; C3 = French-speaking country; C4 = Turkish-speaking country; C5 = English-speaking country; C6 & C7 = Chinese-speaking countries; C8 = Korean-speaking country

and between-country inconsistencies. Concerning Item 5, human scores were affected by how students described the key concept of *increasing heart rate*. In some cases, similar responses received different scores depending on the country. Some human raters marked responses as correct even if they only included numbers indicating elevated heart rates, like 150 or 200, despite the students being asked to provide a brief 'description' of the changes in heart rate. This demonstrates that achieving a perfect agreement between humans and machines is unattainable, especially in multilingual contexts.

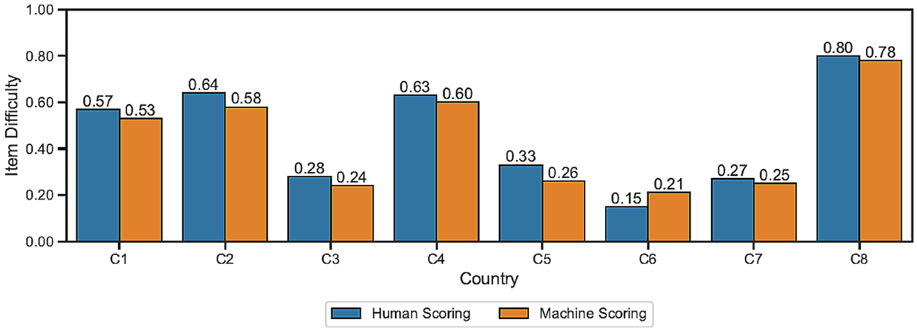

Fig. 4 Country-by-item difficulty of item 4. *Note* C1 & C2 = German-speaking countries; C3 = French-speaking country; C4 = Turkish-speaking country; C5 = English-speaking country; C6 & C7 = Chinese-speaking countries; C8 = Korean-speaking country

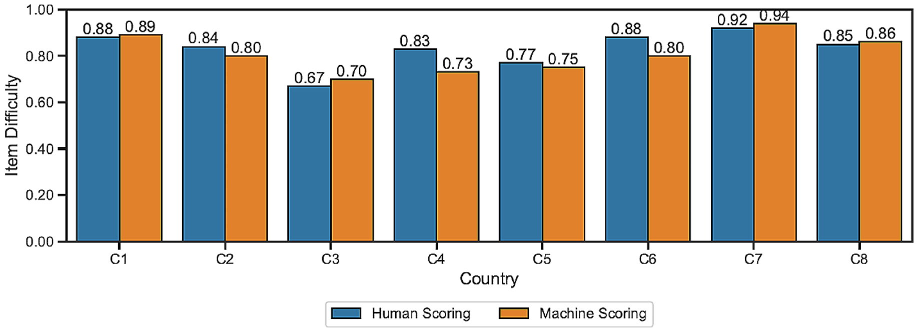

Fig. 5 Country-by-item difficulty of item 5. *Note* C1 & C2 = German-speaking countries; C3 = French-speaking country; C4 = Turkish-speaking country; C5 = English-speaking country; C6 & C7 = Chinese-speaking countries; C8 = Korean-speaking country

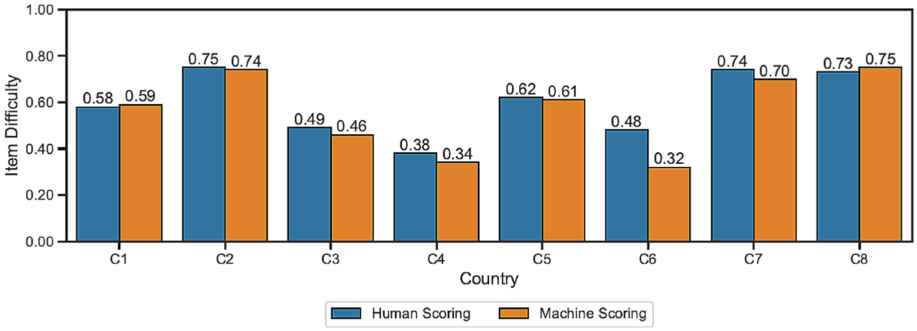

Fig. 6 Country-by-item difficulty of item 6. *Note* C1 & C2 = German-speaking countries; C3 = French-speaking country; C4 = Turkish-speaking country; C5 = English-speaking country; C6 & C7 = Chinese-speaking countries; C8 = Korean-speaking country

Lastly, disparities between human and machine scores may stem from the uncertainty in the true score, defined as the expected value of the observed score. True scores can be uncertain for ambiguous responses, especially for misspelled responses. The level of acceptable misspellings can be subjective and may vary depending on human raters. For example, the keyword of Item 6 was *rust* which is 生銹 (shēngxiù/) in traditional Chinese characters. However, in C6, misspelled or non-existent words were scored as correct responses by the human rater due to their phonetic similarity: (a) 生秀, (b) 生受, (c) 生廋, (d) 生瘦, and (e) 生獸. The misspelled second characters (a) 秀 (/xiù/), (b) 受

(/shòu/), (c) 廋 (/sōu/), (d) 瘦 (/shòu/), and (e)獸 (/shòu/) have the identical or similar pronunciation of the correct character 銹 (/xiù/). These responses, constituting 44% of the misalignments in C6, were scored as incorrect by the machine. This led to a substantial negative SMD (-0.32) and a large disparity in item difficulty between human and machine scores (0.16). Chinese native speakers said that human scores may have differed in whether the raters considered these misspelled responses as correct.

Directions for future research

We observed that the differences between human and machine scores were derived from various factors, not just the error of ANN classifications. While benchmarking human scoring is still important, Bennett and Bejar (1998) stressed that relying solely on human scores to assess automated scoring is counterproductive due to the fallibility of human raters. Rather, the central focus in automated scoring should be on the accuracy, consistency, and fairness of machine scores. Thus, it is imperative to investigate whether machine scores accurately capture the intended construct, evaluate the alignment of features used in automated scoring with the rubric, and identify any potential biases or fairness issues. (Attali, 2013; Bennett & Zhang, 2015; Bowler et al., 2020; Madnani & Cahill, 2018). Through comprehensive evaluation and validation, we can advance toward more reliable and accurate automated scoring.

Limitation

One limitation of this study is the absence of human evaluation of MT quality. Although we generally reviewed MT by comparing text length similarities between the original and translated responses and checking any hallucinations from ChatGPT, we did not use an MT quality metric such as the bilingual evaluation understudy (BLEU) metric - which measures the word-based overlap between MT output and professionally translated human text. However, considering our ultimate goal to expand automated scoring to ILSAs administered in over 100 languages, it is crucial to employ automated MT evaluation rather than relying on human judgment to assess MT quality. While we used a combination of multiple MT and LOR as one approach, future research can explore the integration of automated MT evaluation into automated scoring.

Conclusion

This study investigated the potential of automated scoring in ILSAs. The findings showed that automated scoring with MT could be a promising support or alternative to human scoring, which has inherent concerns of inconsistency and high expense. With the ongoing advancement in MT and ANNs, we anticipate the performance of automated scoring will continue to improve, making it easier to use and reliably score short CR items in ILSAs. We suggest that future research expands the scope of automated scoring to more languages and countries with advanced NLP and ANN approaches.

Received: 4 August 2023 / Accepted: 22 March 2024
[redacted]

References

Abiodun, O. I., Jantan, A., Omolara, A. E., Dada, K. V., Mohamed, N. A., & Arshad, H. (2018). State-of-the-art in artificial neural network applications: A survey. *Heliyon*, *4*(11), e00938. https://doi.org/10.1016/j.heliyon.2018.e00938.

Attali, Y. (2013). Validity and reliability of automated essay scoring. In M. D. Shermis, & J. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 181–198). Routledge.

Balahur, A., & Turchi, M. (2012, July). Multilingual sentiment analysis using machine translation? *Proceedings of the 3rd workshop in computational approaches to subjectivity and sentiment analysis*, 52–60. https://aclanthology.org/W12-3709.pdf.

Bennett, R. E. (1991). On the meanings of constructed response. ETS Research Report Series. https://doi.org/10.1002/j.2333-8504.1991.tb01429.x.

Bennett, R. E., & Bejar, I. I. (1998). Validity and automated scoring: It's not only the scoring. *Educational Measurement: Issues and Practice*, *17*(4), 9–17. https://doi.org/10.1111/j.1745-3992.1998.tb00631.x.

Bennett, R. E., & Zhang, M. (2015). Validity and automated scoring. In F. Drasgow (Ed.), *Technology and testing: Improving educational and psychological measurement* (pp. 142–173). Routledge. https://doi.org/10.4324/9781315871493-8.

Berrar, D. (2019). Cross-validation. *Encyclopedia of Bioinformatics and Computational Biology*, *1*, 542–545. https://doi.org/10.1016/B978-0-12-809633-8.20349-X.

Bowler, E., Fretwell, P. T., French, G., & Mackiewicz, M. (2020). Using deep learning to count albatrosses from space: Assessing results in light of ground truth uncertainty. *Remote Sensing*, *12*(12), 2026. https://doi.org/10.3390/rs12122026.

Britz, D., Goldie, A., Luong T, M., & Le, Q. (2017). Massive exploration of neural machine translation architectures. *arXiv*. https://doi.org/10.48550/arXiv.1703.03906.

Cahill, A., & Evanini, K. (2020). Natural language processing for writing and speaking. In D. Yan, A. A. Rupp, & P. W. Foltz (Eds.), *Handbook of automated scoring: Theory into practice* (pp. 69–92). Chapman and Hall/CRC. https://doi.org/10.1201/9781351264808.

Caswell, I. (2022). Google Translate learns 24 new languages. *Google*. https://blog.google/products/translate/24-new-languages/.

Chen, X., Chen, C., Zhang, D., & Xing, Z. (2019). Sethesaurus: Wordnet in software engineering. *IEEE Transactions on Software Engineering*, *47*(9), 1960–1979. https://doi.org/10.1109/TSE.2019.2940439.

Darling-Hammond, L., & Adamson, F. (2010). *Beyond Basic Skills: The Role of Performance Assessment in Achieving 21st Century Standards of Learning*. Stanford Center for Opportunity Policy in Education. https://globaled.gse.harvard.edu/sites/projects.iq.harvard.edu/files/geii/files/beyond-basic-skills-role-performance-assessment-achieving-21st-century-standards-learning-report_0.pdf.

de Vries, E., Schoonvelde, M., & Schumacher, G. (2018). No longer lost in translation: Evidence that Google translate works for comparative bag-of-words text applications. *Political Analysis*, *26*(4), 417–430. https://doi.org/10.1017/pan.2018.26.

Foltz, P. W., Yan, D., & Rupp, A. A. (2020). The past, present, and future of automated scoring. In D. Yan, A. A. Rupp, & P. W. Foltz (Eds.), *Handbook of automated scoring: Theory into practice* (pp. 1–10). Chapman and Hall/CRC. https://doi.org/10.1201/9781351264808.

Google (2023a). March 28). Language Support. *Google Cloud*https://cloud.google.com/translate/docs/languages.

Google (2023b, March 28). Cloud Translation Pricing. *Google Cloud*. https://cloud.google.com/translate/pricing.

Han, F., Jiang, J., Ling, Q. H., & Su, B. Y. (2019). A survey on metaheuristic optimization for random single-hidden layer feedforward neural network. *Neurocomputing*, *335*, 261–273. https://doi.org/10.1016/j.neucom.2018.07.080.

Hendy, A., Abdelrehim, M., Sharaf, A., Raunak, V., Gabr, M., Matsushita, H., Kim, Y., Afify, M., & Awadalla, H. H. (2023). *How good are gpt models at machine translation? a comprehensive evaluation*. arXiv. https://arxiv.org/abs/2302.09210.

Hovy, D., & Prabhumoye, S. (2021). Five sources of bias in natural language processing. *Language and Linguistics Compass*, *15*(8), e12432. https://doi.org/10.1111/lnc3.12432.

Hutchins, J. (2007). Machine translation: A concise history. *Computer Aided Translation: Theory and Practice*, *13*(29–70), 11.

Jiao, W., Wang, W., Huang, J. T., Wang, X., & Tu, Z. (2023). Is ChatGPT a good translator? A preliminary study. arXiv. https://arxiv.org/pdf/2301.08745.pdf.

Jung, J. Y., Tyack, L., & von Davier, M. (2022). Automated scoring of constructed-response items using artificial neural networks in international large-scale assessment. *Psychological Test and Assessment Modeling*, *64*(4), 471–494. https://www.psychologie-aktuell.com/fileadmin/Redaktion/Journale/ptam_2022-4/PTAM_2022-4_5.pdf.

Kim, Y. (2014). Convolutional neural networks for sentence classification. *arXiv*. https://doi.org/10.48550/arXiv.1408.5882.

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

Le, T. N., & Huynh, H. T. (2016). Liver tumor segmentation from MR images using 3D fast marching algorithm and single hidden layer feedforward neural network. *BioMed Research International*, *2016*. https://doi.org/10.1155/2016/3219068.

Liu, O. L., Brew, C., Blackmore, J., Gerard, L., Madhok, J., & Linn, M. C. (2014). Automated scoring of constructed-response science items: Prospects and obstacles. *Educational Measurement: Issues and Practice*, *33*(2), 19–28. https://doi.org/10.1111/emip.12028.

Lucas, C., Nielsen, R. A., Roberts, M. E., Stewart, B. M., Storer, A., & Tingley, D. (2015). Computer-assisted text analysis for comparative politics. *Political Analysis*, *23*(2), 254–277. https://doi.org/10.1093/pan/mpu019.

Madnani, N., & Cahill, A. (2018, August). Automated scoring: Beyond natural language processing. *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 1099–1109). https://aclanthology.org/C18-1094.

Mallinson, J., Sennrich, R., & Lapata, M. (2017). Paraphrasing revisited with neural machine translation. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, *1*(Long Papers), 881–893. https://aclanthology.org/E17-1083.

Martin, M. O., von Davier, M., & Mullis, I. V. (2020). Methods and Procedures: TIMSS 2019 Technical Report. *International Association for the Evaluation of Educational Achievement*. https://timssandpirls.bc.edu/timss2019/methods/.

McClellan, C. A. (2010). Constructed-Response Scoring—Doing it Right. *ETS R&D Connections*, *13*, 1–7. Princeton, NJ: Educational Testing Service. http://www.ets.org/research/policy_research_reports/rdc-13.

META (2022, July 6). *New AI model translates 200 languages, making technology accessible to more people*https://about.fb.com/news/2022/07/new-meta-ai-model-translates-200-languages-making-technology-more-accessible/.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv*. https://doi.org/10.48550/arXiv.1301.3781.

Montalvo, S., Martínez-Unanue, R., Fresno, V., & Capilla, R. (2015). Multilingual information Access on the web. *Computer*, *48*(7), 73–75. https://doi.org/10.1109/MC.2015.203.

Mullis, I. V. S., & Martin, M. O. (Eds.). (2017). *TIMSS 2019 Assessment Frameworks*. Retrieved from Boston College, TIMSS & PIRLS International Study Center website: http://timssandpirls.bc.edu/timss2019/frameworks/.

Myford, C. M., & Wolfe, E. W. (2009). Monitoring rater performance over time: A framework for detecting differential accuracy and differential scale category use. *Journal of Educational Measurement*, *46*(4), 371–389. https://doi.org/10.1111/j.1745-3984.2009.00088.x.

OpenAI (2023b). Pricing. *OpenAI*. https://openai.com/pricing.

OpenAI (2023a). Models. *OpenAI*. https://platform.openai.com/docs/models/overview.

Page, E. B. (1966). The imminence of… Grading essays by Computer. *The Phi Delta Kappan*, *47*(5), 238–243. https://www.jstor.org/stable/20371545.

Prakash, A., Hasan, S. A., Lee, K., Datla, V., Qadir, A., Liu, J., & Farri, O. (2016). *Neural paraphrase generation with stacked residual LSTM networks*. arXiv. https://doi.org/10.48550/arXiv.1610.03098.

Prasanna, P. L., & Rao, D. R. (2018). Text classification using artificial neural networks. *International Journal of Engineering & Technology*, *7*(1.1), 603–606. https://doi.org/10.14419/ijet.v7i1.1.10785.

Ramineni, C., & Williamson, D. M. (2013). Automated essay scoring: Psychometric guidelines and practices. *Assessing Writing*, *18*(1), 25–39. https://doi.org/10.1016/j.asw.2012.10.004.

Song, G., Ye, Y., Du, X., Huang, X., & Bie, S. (2014). Short text classification: A survey. *Journal of Multimedia*, *9*(5).

Sutskever, I., Martens, J., & Hinton, G. E. (2011). Generating text with recurrent neural networks. *Proceedings of the 28th international conference on machine learning* (*ICML-11*) (pp. 1017–1024). https://www.cs.toronto.edu/~jmartens/docs/RNN_Language.pdf.

Timothy, M. (2023, March 10). *How to Use ChatGPT as a Language Translation Tool*. https://www.makeuseof.com/how-to-translate-with-chatgpt/.

von Davier, M., Tyack, L., & Khorramdel, L. (2023). Scoring graphical responses in TIMSS 2019 using artificial neural networks. *Educational and Psychological Measurement*, *83*(3), 556–585. https://doi.org/10.1177/00131644221098021.

Wang, S., & Jiang, J. (2016). Machine comprehension using match-lstm and answer pointer. *arXiv*. https://doi.org/10.48550/arXiv.1608.07905.

Wang, S., Tu, Z., Tan, Z., Wang, W., Sun, M., & Liu, Y. (2021). *Language models are good translators*. arXiv. https://arxiv.org/pdf/2106.13627.pdf.

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice*, *31*(1), 2–13. https://doi.org/10.1111/j.1745-3992.2011.00223.x.

Wolfe, E. W., & McVay, A. (2012). Application of latent trait models to identifying substantively interesting raters. *Educational Measurement: Issues and Practice*, *31*(3), 31–37. https://doi.org/10.1111/j.1745-3992.2012.00241.x.

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. https://doi.org/10.1002/ets2.12169.

Zhai, X., Yin, Y., Pellegrino, J. W., Haudek, K. C., & Shi, L. (2020). Applying machine learning in science assessment: A systematic review. *Studies in Science Education*, *56*(1), 111–151. https://doi.org/10.1080/03057267.2020.1735757.

Zhang, M. (2013). Contrasting automated and human scoring of essays. *ETS R&D Connections*, *21*(2), 1–11. https://www.ets.org/Media/Research/pdf/RD_Connections_21.pdf.

Publisher's Note