Boston College

Lynch School of Education and Human Development

Department of

Measurement, Evaluation, Statistics, and Assessment

AUTOMATED SCORING IN INTERNATIONAL LARGE-SCALE ASSESSMENTS: FEASIBILITY, MULTILINGUAL COMPARABILITY, AND SCALABILITY

Dissertation

by

JI YOON JUNG

submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

July 2024

ABSTRACT

AUTOMATED SCORING IN INTERNATIONAL LARGE-SCALE ASSESSMENTS: FEASIBILITY, MULTILINGUAL COMPARABILITY, AND SCALABILITY

Ji Yoon Jung, Author

Matthias von Davier, Chair

Automated scoring has received considerable attention in educational measurement, even before the era of artificial intelligence. However, its application to constructed response (CR) items in international large-scale assessments (ILSAs) remains largely underexplored due to the complexity of tackling multilingual responses spanning often over 100 different language versions. This doctoral dissertation aims to address this issue by progressively expanding the scope of automated scoring from several countries in TIMSS 2019 to all participating countries in TIMSS 2023. We delved into the feasibility of automated scoring across diverse linguistic landscapes, encompassing high-resource and low-resource languages. We examined two machine learning methodologies—supervised and unsupervised learning—integrating them with cutting-edge machine translation techniques. Our findings demonstrated that automated scoring can serve as a reliable and cost-effective measure for quality assurance in ILSAs, significantly reducing the reliance on secondary human raters. Ultimately, the adoption of automated scoring instead of human scoring in the foreseeable future will promote the broader use of innovative open-item formats in ILSAs.

ACKNOWLEDGEMENTS

I am deeply grateful to my academic advisor, Dr. Matthias von Davier, for his professional guidance and encouragement throughout my four-year doctoral journey. His expertise and insightful advice have consistently inspired me to move forward and do my best work. Thanks to his mentorship, guidance, and kindness, I have happily learned and grown as a self-motivated researcher.

I would like to acknowledge my committee members, Dr. Henry Braun and Dr. Brian Smith, for their ongoing support and feedback throughout each stage of the process, including the preproposal, proposal, and defense. Their critical feedback has provided me with valuable insights that have contributed to the completion of this dissertation.

I have to express my heartfelt thanks to my colleagues at the TIMSS & PIRLS International Study Center, especially Dr. Lale Khorramdel, Dr. Ummugul Bezirhan, and Lillian Tyack. Working as a research assistant at the Study Center was an invaluable experience, during which I learned a lot about data analysis, methodologies, implementation, and educational measurement itself. Also, I want to thank my friends in the MESA program, particularly Echo, Dihao, and Clara, for their friendship and support.

I wish to express my deepest gratitude to my family - my husband, Ikbeom, and my daughter, Evelyn (Sunyu). Thanks to his unwavering love and support, I have truly enjoyed my academic journey. Starting my academic career at Purdue with his encouragement was one of the wonderful decisions I ever made. I also thank my beautiful daughter, Evelyn, for her pure love and joy. I am also grateful to my parents for their unconditional love and support, which I felt deeply despite the distance between Korea and the US.

TABLE OF CONTENTS

CHAPTER 1. INTRODUCTION

This doctoral dissertation is built around three research studies described in a set of coordinated but separately published research articles (see Table 1 below). The unifying theme of the three papers is the exploration of how artificial intelligence (AI) can be leveraged for automated scoring of constructed response (CR) items in international large-scale assessments (ILSAs).

The first paper examines the potential of automated scoring using short CR items derived from the TIMSS 2019 database. It also explores if the adoption of Item Response Theory (IRT) to produce expected responses can contribute to cleaning data by removing potentially inconsistent or incorrect human scores.

The second paper expands the scope of the first paper to a multilingual context and studies the feasibility of automated scoring across languages as an intermediate step to the operational use of the approach in ILSAs. This paper 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 apply AI-based automated scoring to multilingual student responses from eight countries and six different languages, using six CR items from TIMSS 2019.

The third paper focuses on the scalability of automated scoring combined with machine translation to all participating countries and languages in the TIMSS 2023. This study expands the earlier research studies and focuses on the comparability of automated scoring to human scoring across countries and languages. Considering the significant advances in machine learning made by large language models (LLMs), initiated by the global release of OpenAI's Generative Pre-trained Transformer (GPT) models, we investigated supervised and unsupervised learning approaches. Moreover, this paper examines the potential of automated scoring as a measure for ensuring quality assurance in ILSAs.

In conclusion, the goal of this dissertation is to contribute to establishing the use of multilingual automated scoring in the context of ILSAs.

Table 1 Overview of Papers 1, 2, and 3

		Paper 1	Paper 2	Paper 3
Title		Automated Scoring of Constructed-Response Items Using Artificial Neural Networks in International Large-scale Assessment	Combining Machine Translation and Automated Scoring in International Large-Scale Assessments	Towards the Implementation of Automated Scoring in International Large-scale Assessments: Scalability and Quality Control
Research Questions		● How automated scoring can be applied in ILSAs? ● Does Item Response Theory (IRT) contribute to purifying the data for quality control?	● Can automated scoring achieve comparable performance to human scoring across different countries and languages without compromising the psychometric properties of items? ● Does MT appropriately convert non-English language responses into English to construct a unified cross-lingual automated scoring model? ● What are the sources of misalignment between human and automated scoring?	● Does automated scoring exhibit consistent and accurate performance across all participating countries and languages in TIMSS 2023? ● How can MT tools be effectively used to expand the scope of automated scoring? ● What causes discrepancies between human and automated scoring in a multilingual context?
Data	Cycle	TIMSS 2019	TIMSS 2019	TIMSS 2023
	Grade	Grade 4	Grade 4	Grade 4 (8)
	Country	United States (1 country)	Austria, Germany, France, Hong Kong, Korea, Turkey, Chinese Taipei, and United States (8 countries)	All participating countries (52 countries for items 1 and 2, and 40 countries for item 3)
	Language	English (1 language)	Chinese, English, French, German, Korean, and Turkish (6 languages)	42 languages
	Item	4 short CR items	6 short CR items	3 CR items
Methods		● Item Response Theory ● Bag-of-Words ● Feed-forward Neural Networks	● Multiple Machine Translation ● Bag-of-Words ● Feed-forward Neural Networks	● Machine Translation ● OpenAI's GPT models ● Feed-forward Neural Networks
Status		Published in Psychological Test and Assessment Modeling	Published in Large-scale Assessment in Education	Initial manuscript ready

Paper 1

Automated Scoring of Constructed-Response Items Using Artificial Neural Networks in International Large-scale Assessment

Published as:

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.

Summary

As a starting point, this paper explores the feasibility of automated scoring in ILSAs using the TIMSS 2019 database. The purpose of this paper is twofold: (1) to provide a feasibility study of automated scoring in the monolingual context (i.e., English), and (2) to examine whether IRT can help provide expected responses for purifying the training data. The paper used a conventional natural language processing (NLP) technique (i.e., Bag-of-Words) and feed-forward neural networks (i.e., FNNs) to explore the viability of automated scoring. BoW and FNNs can be efficient strategies to score short constructed response items due to their simplicity and interpretability. The results demonstrated the potential of automated scoring with a high agreement (r=0.91) between human and machine-generated scores. Also, adopting IRT-based scores could be a promising strategy to improve the accuracy of automated scoring since FNNs provide comparable or even slightly higher accuracy when trained on the data filtered by IRT-generated scores.

Paper 2

Combining Machine Translation and Automated Scoring in International Large-Scale Assessments

Published as:

Jung, J.Y., Tyack, L. & von Davier, M. Combining machine translation and automated scoring in international large-scale assessments. *Large-scale Assessments in Education*, *12*, 10 (2024). https://doi.org/10.1186/s40536-024-00199-7

Summary

This paper examines the feasibility of multilingual automated scoring combined with machine translation, using the TIMSS 2019 database. Four alphabetical (i.e., English, French, German, and Turkish) and two non-alphabetical languages (i.e., Chinese and Korean) were selected to test the performance of automated scoring in the multilingual context. The goals of this study are (1) to examine the comparability of automated scoring to human scoring across different countries and languages, (2) to evaluate whether machine translation appropriately converts responses to support building a joint training and validation database, and (3) to identify the sources of misalignment between human and automated scoring. The results show that automated scoring displayed comparable performance to human scoring, especially when the ANNs were trained and tested on responses translated by multilingual LLMs. Furthermore, psychometric characteristics derived from machine scores generally exhibited similarity to those obtained from human scores. Also, we learned that the misclassified responses do not necessarily indicate the error of automated scoring, but may be the sign of human scoring errors and

inconsistencies on occasion. These results can be considered as supportive evidence for the validity of automated scoring for survey assessments.

Paper 3

Towards the Implementation of Automated Scoring in International Large-scale Assessments: Scalability and Quality Control

Summary

This paper explores the generalizability of multilingual automated scoring using participating countries in TIMSS 2023. Despite the considerable attention to automated scoring, international large-scale assessments (ILSAs) have remained a challenge, largely due to the difficulties of scoring multilingual responses. This paper addresses this challenge by investigating two machine learning approaches — supervised and unsupervised learning — for scoring multilingual responses. Across all participating countries and 42 different languages, we examine multiple scoring methods to assess three science CR items in TIMSS 2023. The results showed that the supervised learning approach, particularly combining multiple machine translations with artificial neural networks (MMT_ANNs), outperformed alternative scoring methods. This remarkable performance was attributed to MMT_ANNs providing more suitable translations at both individual response and language levels. Furthermore, unlike human scoring, MMT_ANNs consistently generated precise scores for identical responses within and across countries. These findings indicate the potential of automated scoring as an accurate and cost-effective measure for quality assurance in ILSAs.

Summary

The three papers on automated scoring provide a succession of increasingly more capable solutions for the problem of multilingual scoring of open responses using artificial neural networks. Unlike automated scoring in a single, well-supported language, the methods employed in this dissertation aim to provide viable solutions even for low-resource languages for which elaborate NLP methods are unavailable. The proposed solutions generate a workflow that is generally applicable but also includes checks for translation quality and scoring validity in terms of agreement with expert evaluations, human scorers, and scores derived from machine-scored parts of the assessments. Checking agreement with derived achievement scores and ratings provided by human scorers and assessment development experts ensures that the proposed methods work as well as human ratings or even result in automated scores that are more reliable than human scores and provide improved psychometric properties for subsequent analysis.

CHAPTER 2. PAPER 1

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.

Automated Scoring of Constructed-Response Items Using Artificial Neural Networks in International Large-scale Assessment

Abstract

Although constructed-response items have proven effective in assessing students' higher-order cognitive skills, their wider use has been limited in international large-scale assessments (ILSAs) due to the resource-intensive nature and the challenges associated with human scoring. This study presents automated scoring based on artificial neural networks (ANNs) as feasible support for or as an alternative to, human scoring. We examined the comparability of human and automated scoring for short constructed-response items from TIMSS 2019. The results showed that human and automated scores were highly correlated on average (r=0.91). Moreover, this study found that a novel approach of adopting expected scores generated from item response theory (IRT) can be useful for quality control. The ANN-based automated scoring provided equally high or even improved agreements when it was trained on the data which is weighted or filtered based on IRT-based scores. This study argues that automated scoring has great potential to enable resource-efficient and consistent scoring in place of human scoring and, consequently, facilitate the greater use of constructed-response items in ILSAs.

Introduction

The move to computer-based assessment has enabled international large-scale assessments (ILSAs) to enhance the measurement of student achievement through novel items. Innovative item formats, such as integrated scenarios and tasks that require higher-order cognitive processes frequently include constructed-responses (CR) items. The TIMSS 2019 (Trends in International Mathematics and Science Study) marked the transition to the eTIMSS digital format, incorporating innovative CR items (Martin et al., 2020). Traditional multiple-choice items are thought of as limited to less complex processes such as memorization of key concepts, while CR items are thought to elicit students' deeper understanding by asking them to apply their knowledge in subject areas (Harris et al., 2019; Liu et al., 2014; Maestrales et al., 2021). Unfortunately, CR items have been restrictively used in ILSAs because of the high cost of human scoring: Training human raters to attain the preferred range of agreement is labor-intensive (Braun et al., 1990) and particularly challenging in assessments administered in up to 100 or more language versions. Zhang (2013) stated that the increasing use of CR items in ILSAs is time-consuming and resource-intensive to score due to the high volumes of student responses. Automated scoring holds great potential to enable increased use of CR items, facilitating cost-efficient, fast, and consistent measurement.

There have been many approaches to adopting automated scoring of CR items in large-scale assessments (Braun et al., 1990; Ha & Nehm, 2016; Liu et al., 2014; Liu & Kunnan, 2016; Madnani et al., 2013). These approaches fall into two main categories: 1) handcrafted features-based models and 2) artificial neural network (ANN) based models (Hussein et al., 2019). One example of the former models is the *c-rater* developed by Educational Testing Service (Sukkarieh & Stoyanchev, 2009), which applies scoring rules built from a set of correct model answers with predefined concepts. In contrast, ANN-based scoring models automatically extract the scoring features using machine learning and neural networks (Hussein et al., 2019). For instance, *c-rater-ML* is the automated scoring tool implementing support vector regression. It constructs a statistical model for learning from a set of previously human-scored responses rather than relying on descriptions of the key concepts (Liu & Kunnan, 2016).

Although automated scoring in educational measurement is not new, the adoption of recently developed deep learning techniques for ANNs is lacking in ILSAs. ANNs and natural language processing (NLP) have significantly improved in the past few years (Kersting et al., 2014; Sorin et al., 2020). Recent ANNs have been utilized for automated item generation (von Davier, 2018), automated scoring of graphical input, and complex classification (von Davier et al., 2022). However, less research has investigated the feasibility of automated scoring of CR items in multilingual international assessments possibly due to the challenges associated with machine translation and translation quality control. The present study explores automated scoring of selected CR items from the TIMSS 2019 assessment using ANNs, comparing human rater- and computer-generated scores. This study shows promise for automated scoring in ILSAs, demonstrating how ANN classifiers could be utilized to score CR items in multilingual contexts.

Background

Constructed-Response Items in ILSAs

Large-scale assessments have paid increasingly more attention to technology-enhanced, interactive, and open-response items while shifting from paper-based to computer-based assessments. Technology-based assessment enables the use of more complex and innovative items that generally depend on intricate computer functionality (Bryant, 2017). Particularly, computer-based assessments allow for wider use of CR items. Well-crafted CR items are commonly believed to assess a broader range of higher-order thinking skills (e.g., analyzing, designing, and integrating) in contrast to selected response multiple-choice (MC) items (Darling-Hammond & Adamson, 2010; Hancock, 1994; McClellan, 2010; Jodoin, 2003). CR items may elicit constructive cognitive processes by requiring students to produce their own answers, employing their knowledge and reasoning abilities (Lissitz et al., 2012), while MC items mostly focus on skills such as recognition, recall, or prompted information retrieval (Darling-Hammond & Adamson, 2010). Moreover, CR items may provide deeper insight into student thinking since they allow students to construct heterogeneous or even idiosyncratic answers rather than choosing from a set of responses provided on the test (Federer et al., 2015).

Despite the potential strengths of CR items, their wider use has been limited in ILSAs due to their scoring requirements. The human scoring of CR items is by its nature labor-intensive, costly, and time-consuming, and may lead to validity and reliability issues originating from rater effects such as severity and leniency, inconsistency, and halo effects, as well as other issues (McClellan, 2010; O'Leary et al., 2018; Wahlen et al., 2020; Zhang, 2013). CR items may be prone to problems of scoring subjectivity, especially when scorers are insufficiently trained and human judgment is involved in deciding whether an answer is correct (Brown & Hudson, 1998). Bejar (2012) stated that individual raters build their own mental scoring rubric that can be affected by a variety of factors such as personal attributes or background. These differences in personal mental rubrics may make the scoring behavior of human raters inconsistent and can cause rater effects, resulting in systematic differences in scores (i.e., construct-irrelevant variance). Even after rigorous training and calibration, rater scoring performance cannot be taken for granted (McClellan, 2010), and additional quality control is required to ensure valid inferences from scores. TIMSS employs elaborate scoring consistency checks to mitigate these risks, but these are costly and time-consuming for participating countries.

Fortunately, a growing number of studies have shown that automated scoring can play a viable role in the scoring of CR items, suggesting that high levels of agreement between human rater- and computer-generated scores can be achieved (Ha, 2016; Kersting et al., 2014; Liu et al., 2016; Shermis et al., 2010; Shermis & Burstein, 2013). Automated scoring can be beneficial either by performing second scoring or by substituting for human raters entirely (von Davier et al., 2022). In particular, it not only greatly reduces the cost and time involved in scoring but also provides high consistency and quick score turnaround, offering instant feedback to students (Attali et al., 2008; Higgins et al., 2011; Williamson et al., 1999, Zhang, 2013). Noteworthily, the Duolingo English Test provides experimental evidence of the operational use of automated scoring. Being a computer-adaptive English proficiency test, the Duolingo English Test creates, scores, and analyzes items using machine learning and NLP (Settles et al., 2020). The automated scoring of the Duolingo English Test was found to be highly reliable as can be seen in the moderate-to-high correlations between its computer-generated scores and relevant test scores such as TOEFL writing and IELTS writing (Cardwell et al., 2021).

Progress in Automated Scoring

A number of studies have been conducted to measure the accuracy and reliability of automated scoring of students' written responses (Dikli, 2006; Wahlen et al., 2020). In 1965, Page developed the first automated scoring engine, *Project Essay Grader* (PEG), suggesting the comparability of human scoring and computer scoring (Page, 1966). *PEG* focused on extracting text surface features to predict scores using multiple regression. He analyzed a set of 138 English essays written by high school students in grades 8-12, scoring with four human raters and one computer rater. He not only found that computer-generated scores were similar to human raters ($r = 0.50$) but also asserted that computers will perform better than human raters as individual random errors are eventually eliminated from computers while these have to be assumed in human raters. *E-rater* developed by ETS used surface features like PEG but also considered textual coherence to predict human holistic scores (Enright & Quinlan, 2010; Miller, 2003). *E-rater* provided evidence for construct validity demonstrating that *e-rater* and human raters assess essentially the same construct (Attali, 2007). Although initial findings were encouraging, both PEG and *e-rater* were criticized for their lack of consideration of content or deeper semantic information (Wang, 2020).

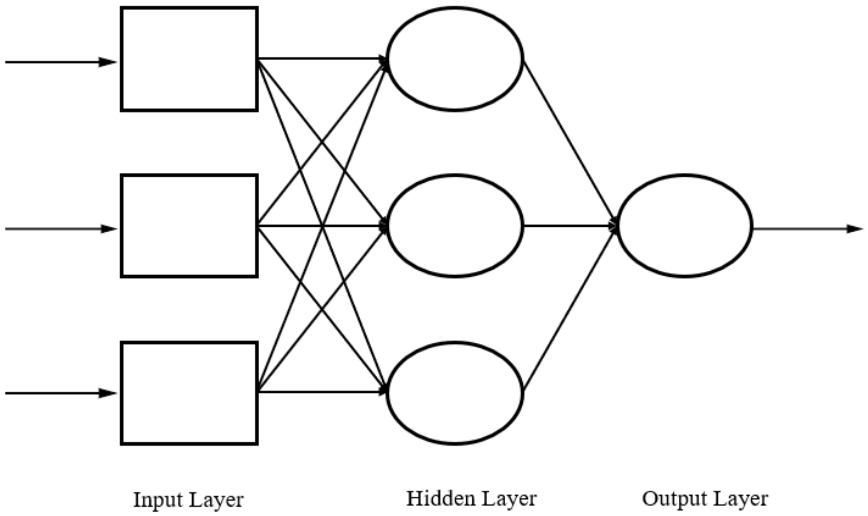
Beyond surface-level models, latent semantic analysis (LSA) is a machine learning-based technique that uncovers the underlying semantic structure using a singular value decomposition (Landauer et al., 1998; Landauer & Dumais, 1997). LSA deduces the relationship between words and documents, aiming to quantify the deeper semantic content (Hearst, 2000). There have been consistent improvements in LSA and LSA-based approaches are still being employed for automated scoring. Using generalized LSA, Islam, and Hoque (2010) trained on 960 essays written by undergraduate students and, subsequently, analyzed 120 testing essays. They achieved high accuracy of automated scoring with human-machine score correlations ranging from 0.89 to 0.96. With an LSA-based automated scoring, LaVoie et al. (2020) scored short answer responses ($N = 1{,}863$) written by Reserve Officers' Training Corps cadets from the Consequences Test, a measure of creativity and divergent thinking. Automated scores demonstrated very high convergence with human raters ($r = 0.94$) and provided similar patterns of predictive and concurrent validity as human scores (i.e., scores from Cadet Order of Merit Listing and Cadet Grade Point Average).

Rapid advances in machine learning enable automated scoring to be more adaptable and accurate than traditional approaches. Artificial neural networks (ANNs), and especially deep neural networks, are powerful machine learning algorithms that simulate the information processing capability of the human brain (Dongare et al., 2012; Williamson et al., 2004). Many ANNs consist of three types of layers such as an input layer of neurons, one or more hidden layers, and a final layer of output neurons (Wang, 2003). The current study focuses on feed-forward neural networks with a single hidden layer (see Figure 1) where the inputs are fed into the input layer without any feedback from the output layer. A hidden layer exists in-between input and output layers and higher-order statistics are extracted to generate output layers (Sazli, 2006).

Through repeated exposure to data (input and desired output), ANNs learn from the data by conditioning individual neurons either excitatory or inhibitory to certain patterns. The power of ANNs is that they can be applied to new data once they learn patterns and relationships in the data (Agatonovic-Kustrin & Beresford, 2000; Wesolowski & Suchacz, 2012). Nowadays, ANNs are widely used for a variety of purposes including classification, prediction, pattern recognition, or clustering (Abiodun et al., 2018), and more recently, natural language generation (NLG; e.g., Karpathy, 2015; Vaswani et al., 2017; von Davier, 2019).

Figure 1

Single Hidden Layer Neural Networks



The latest advancements in natural language processing (NLP) are also playing an important role in education including automated scoring, automated item generation, writing assistants, and automated feedback (Alhawiti, 2014; Flor & Hao, 2021, Lee et al., 2019). NLP aims to program machines to process spoken or written language (natural language) input and turn it into a useful form of representation (Chary et al., 2019; Rokade et al., 2018). In automated scoring, computers are trained to learn the relationship between features of student responses (e.g., number of words, instances of conjunctions) and human-generated scores (Correnti et al., 2020). After forming these associations, features of new student responses are evaluated with machine learning algorithms, and then computers produce predicted scores for individual responses. Recent neural networks can be effectively trained in solving NLP tasks by addressing many challenges accompanied by the processing of natural languages, such as breaking sentences, extracting semantic information, converting unstructured data into a structured format, or translating multilingual data (Bahja, 2020).

Despite the huge promise of automated scoring based on ANNs and NLP, little is known about its application to multilingual international assessment. The current study aimed to apply automated scoring and examine its comparability with human scores in the context of ILSAs. We implemented supervised learning algorithms for ANNs on constructed-response items from TIMSS 2019.

Methods

Item Selection and Rationale

This present study used four released CR items from TIMSS 2019 and analyzed student responses collected from the United States. The current work describes the methods and results for US English responses. The technologies used were selected with an eye to multilingual capabilities and generalizability to languages other than English (results for other languages are reported in a separate paper). All four items were dichotomously scored items in which students received full credit for correct responses and no credit for incorrect responses. The four items were homogenous in terms of eliciting a short response from students. Two items (SE71054 & SE71077) were relatively easy, while the other two items (ME72209 & SE62005) were moderate-to-high difficulty. The sample size of each item was 1,230 (SE71054), 1,238 (SE71077), 1,197 (ME72209), and 1,239 (SE62005) students.

These items were selected since we wanted to examine whether automated scoring using ANNs can produce computer-generated scores that are comparable to human-generated scores for short CR items in TIMSS. The average lengths of responses for SE71054, SE71077, ME72209, and SE62005 were 59 words, 63 words, 99 words, and 114 words, respectively. Also, according to human scores, 62.0% and 58.3% of students provided correct responses for SE71054 and SE71077, respectively. In contrast, merely 18.3% and 29.2% of students wrote correct responses for ME72209 and SE62005, respectively. The human scores were obtained from professional human raters who scored the responses based on detailed scoring guides after receiving extensive training by the TIMSS & PIRLS International Study Center (Fishbein et al., 2020).

Procedure for Automated Scoring

Preparing Data Set

Using simple holdout validation, the data was split into training and validation sets at a ratio of 8:2; student responses were randomly assigned to the training (80%) and validation set (20%). The holdout method was introduced to avoid overfitting often caused by training and evaluating a model on the same data (Raschka, 2018). Both training and validation set preserved the same class distribution of the data since the random sampling occurred within each class (correct vs. incorrect responses). Also, a single unweighted and unfiltered validation set was used for individual items to evaluate the performance of ANNs across different training approaches.

Preprocessing

Preprocessing is an essential component of text classification since it converts the original form of natural language into a more suitable form to process (Romanov et al., 2019). In this study, student responses in the training set were preprocessed in multiple steps using NLP tools (e.g., *quanteda, quanteda.textstats* & *hunspell*) available in R: tokenization, lowercasing, spelling correction, stopwords removal, and stemming. (Benoit et al., 2018; Benoit et al., 2021; Ooms, 2019).

Step 1: Tokenization

Tokenization is the process of splitting a stream of written text into individual words, phrases, or other meaningful elements called tokens (Uysal & Gunal, 2014), making it easy to manage text data with a set of tokens. In this study, punctuations were replaced with a single whitespace and then student responses were tokenized into words. For example, the sentence "*Whales are mammals.*" was converted to "*Whales are mammals*" without a period, and then was split into three tokens of "*Whales*", "*are*", and "*mammals*".

Step 2: Lowercasing

Lowercasing refers to the conversion of every word in the data to lowercase so that semantically identical words (e.g., "*Whales*" and "*whales*") would not be regarded as different tokens (Oliinyk et al., 2020). It is helpful to increase the quality of classification in terms of accuracy and dimension reduction disregarding domain and language (Uysal & Gunal, 2014). After lowercasing, the aforementioned tokens were transformed to "*whales*", "*are*", "*mammals*".

Step 3: Spelling Correction

As students should not be penalized for their spelling errors (Madnani et al., 2013), we implemented a unique spelling correction method incorporating edit distance and *hunspell* package (Ooms, 2019) in R. First, separate lists of correctly spelled words (i.e., good words) and misspelled words (i.e., bad words) were created from the training set. To further the example above "whales" would be a member of the list of good words, while "whalkes" would be a bad word, as it is not a correctly spelled word found in customarily used spelling correction dictionaries. Next, two different suggested word lists for bad words were generated; one was based on the edit distance approach while the other was on the *hunspell* dictionary. Here, edit distance (d) denotes the minimum number of operations (e.g., insertions, deletions, and replacements) needed to transform a bad word (s_i) into a good word (s_j). The final good word will be chosen among a list of good words where l is the total number of suggested good words. In the example, the edit distance of the bad word "whalkes" relative to the good word "whales" is one, as only a single deletion of the letter “k” is needed. For the edit distance-based list, a good word showing the maximum value from the following equation was selected as an alternative for individual bad words.

$$good\,word(i,j) = \; max_{(j=1\ldots l)}\left[\frac{log(frequency\; of s_j)}{d(s_i,s_j)^2}\right] \tag{1}$$

In other words, we selected the final good word to replace a bad word in the training data based on the (log) word frequency of the good word, weighted by the inverse of the squared edit distance between a good word and a bad word.

For the *hunspell*-based list, the most frequently appearing word in the training set was chosen as an alternative among the suggested words offered by the *hunspell* dictionary. Next, the final suggested word list was produced by comparing the edit distance list and hunspell list; a good word from the edit distance list was used if the edit distance between the good word and the bad word was less than 3, otherwise, a bad word was replaced with a good word from the *hunspell*-based list.

This procedure ensured that any incorrectly spelled word (bad word) in the training set was corrected predominantly based on correctly spelled words (good words) in the remainder of the training set. Only if the edit distance to any good word exceeded a certain threshold, other (*hunspell*) suggested words were used for spelling correction.

It is important to note that the list of good words comprises all correctly spelled words in the training set, irrespective of whether the response containing the words was scored correctly or incorrectly.

Step 4: Stopwords Removal

Stopwords (e.g., *so*, *the*, *from*) are frequently occurring words that barely deliver any information (Ghag & Shah, 2015). For instance, “*are*” from the aforementioned three tokens were removed, and thereby, “*whales*” and “*mammals*” remained.

Step 5: Stemming

Stemming is the process of reducing words to their word roots (i.e., stem) generally done by deleting any attached suffixes or prefixes from the word (Jivani, 2011). Stemming converted “*whales*” and “*mammals*” into “*whale*” and “*mammal*”, respectively.

Bag-of-Words

After preprocessing, the bag-of-words model was applied to represent student responses with a vector of word counts that occur in them (Boulis & Ostendorf, 2005). This vectorized representation of words (i.e., features) enables machines to process the features for training and classification (Shao et al., 2018). In this study, only features (words) appearing at least 0.05% in the training set were included in the feature matrix for more efficient dimension reduction.

Training and Testing the Model

All models were trained using ANNs with the *caret* package in R (Kuhn et al., 2020). The ANNs used in this study were fully-connected feed-forward neural networks, consisting of three layers (i.e., one input layer, one hidden layer, and one output layer). The number of neurons in the input layer was equal to the number of features extracted from the bag of words. The two hyper-parameters in the hidden layer (e.g., size and decay) were optimized for the best candidate model after multiple iterations. The output layer was one neuron, indicating either a correct or an incorrect response.

5-fold cross-validation (CV) was implemented on a training set (80%) and the final model was tested on a previously unseen validation set (20%) to avoid potential data leakage in preprocessing; in spelling correction, the lists of good words and bad words were created based on the full training set, and then spelling correction was performed for the test set, therefore an independent unseen validation dataset was withheld which was not used in any preprocessing The presence of an independent validation set prevents possible data leakage from the training set to the validation set and enables a more appropriate evaluation of the final model performance.

Regarding the validation set, the same preprocessing procedure was applied as the training set. The only difference was that bad words in the validation set were replaced with good words in the suggested words list created from the training set. The preprocessed validation set was represented on the feature matrix extracted from the training set so the models classified the validation set using the same feature matrix.

Different Approaches for Data based on IRT-based Scores

This study used three different approaches for weighting the training data to investigate whether data manipulation has any impact on the classification performance of models: 1) all data unweighted, 2) all data weighted, and 3) match data unweighted. *All data unweighted* was untouched raw data while *all data weighted* and *match data unweighted* were based upon the agreement between scores generated by human raters and item response theory (IRT; Lord & Novick, 1968). As some human raters produce incorrect or inconsistent scores (von Davier et al., 2022), this study used the scores generated from IRT scaling and population modeling as a second opinion to purify the data. Using additional IRT-based scores can be helpful to obtain truly correct or incorrect responses by mitigating the inconsistencies of human scoring. Given that the quality of the training set influences the accuracy and efficiency of machine learning tasks (Gupta et al., 2021), having additional expected scoring allows for obtaining cleaner data.

Specifically, the item parameters (i.e., item discrimination and difficulty) reported in the eTIMSS 2019 (Foy et al., 2020, Chapter 12) were fixed in a 2-parameter logistic (2PL) IRT model (see Table 1) to calculate the probability of a student *n* with the ability θ to get the correct response for an item *i*. Item discrimination (*a*) is the point biserial correlation between a correct response to the item and the total score. Item difficulty (*b*) is the average percentage of students who correctly responded to the item (Martin et al., 2017). With population modeling, the general student proficiency (*θ*) was computed by considering the relation between student proficiency and contextual variables (von Davier, 2020, Chapter 11).

$$P_{i,n}(\theta) = \frac{exp[a_i(\theta_n - b_i)]}{1+exp[a_i(\theta_n - b_i)]} \tag{2}$$

Table 1

The IRT Item Parameters

Item	*a* (discrimination)	*b* (difficulty)
SE71054	0.941	0.272
SE71077	1.100	0.285

ME72209	1.057	1.470
SE62005	1.250	0.666

Next, the IRT-based scores were generated with a maximum a priori (MAP) estimation which indicates the highest probability for student *n* to solve an item *i*. This estimation allows for the comparison of the human-generated score $x_{i,n[r]}$ by rater *r* and IRT-based score $y_{i,n[max]}$. If MAP is above 0.5, 1 was assigned as the IRT-expected score, otherwise, 0 was assigned. Human-generated scores can either agree upon or disagree with IRT-based scores. For instance, if the human-generated score and IRT-based score are both either 1 or both 0 for student *n*'s response to item *i*, we can say the human score and IRT-based score are matched.

$$MAP = y_{i,n[max]} = max_{(x=0,1)}\{P(X = x|\theta_n, a_i, b_i)\} \tag{3}$$

All data weighted included all student responses regardless of the match between the human-based scores $x_{in[r]}$ and the IRT-based scores $y_{in[max]}$. After holding out the 20% of student responses from the whole dataset for validation, the training set consisted of the matching and mismatching responses at a weight ratio of 2:1; the responses where the human and IRT-based scores matched included both human and IRT ratings, so they were effectively doubled while for the responses for which human and IRT scores did not match we only used the human ratings in the training set. The 2:1 ratio was determined to emphasize the responses where the human and IRT-base scores agree upon, with the assumption that human scores for the matching responses are more reliable than for the mismatching responses. Therefore, the existence of IRT scores can be regarded as similar to a second scorer's evaluation for the matching responses. Concerning *match data unweighted*, this data only consisted of student responses for which human and IRT scores agreed upon.

Results

Sample Sizes for Different Data based on IRT-based Scores

The sample sizes of *all data unweighted*, *all data weighted,* and *match data unweighted* can be found in Table 2. On average, 78% of IRT-based scores matched the human-generated scores; 72%, 79%, 83%, and 78% of matches were found for SE71054, SE71077, ME72209, and SE62005, respectively.

Table 2

Sample Sizes for Different Approaches for Data based on IRT-based Scores

Item	Train			Validation
	All data unweighted	All data weighted	Match data unweighted	
SE71054	985	1694	709	245
SE71077	991	1788	797	247
ME72209	958	1756	798	239
SE62005	992	1776	784	247

* *Note.* Match: human score = IRT-based score; all data weighted: match: mismatch = 2:1

Notably, filtering the data based on a match between the human and IRT-based scores did not harm the representativeness of the raw data. Table 3 showed that the class distribution in *all data unweighted* was maintained in *match data unweighted* for most items (SE71054, SE7107, and SE62005). The only exception was ME72209, which showed an imbalance in *match data* *unweighted*. The ratio of incorrect and correct responses for ME72209 changed from 79.9%:20.1% in *all data unweighted* to 93.6%:6.4% in *match data unweighted*. The IRT model may overpredict incorrect responses for this item because of its high level of difficulty.

Table 3. Class Distribution of All Data Unweighted and Match Data Unweighted

Item	Difficulty (*p*)	All data unweighted			Match data unweighted		
		Incorrect	Correct	Sample Size	Incorrect	Correct	Sample Size
SE71054	0.63	459 (37.3%)	771 (62.7%)	1230	327 (36.9%)	558 (63.1%)	885
SE71077	0.57	528 (42.6%)	710 (57.4%)	1238	409 (42.1%)	563 (57.9%)	972
ME72209	0.20	957 (79.9%)	240 (20.1%)	1197	933 (93.6%)	64 (6.4%)	997
SE62005	0.30	867 (70%)	372 (30.0%)	1239	666 (68.9%)	300 (31.1%)	966

* *Note.* Match: human score = IRT-based score

Performance of Automated Scoring Using ANNs

The automated scoring using ANNs was evaluated in comparison to human-generated scores. First, the performance of the automated scoring was comparable to human scoring across all four items. (see Table 4). For easy items (SE71054 & SE71077), a substantial agreement was found across all approaches to data; $0.93 \leq r \leq 0.94$ in all data unweighted, 0.92 0.94 in all data weighted, and $0.93 \leq r \leq 0.96$ in match data unweighted. The relatively difficult items (ME72209 & SE62005) also showed very high agreement for all approaches to data; $0.85 \leq r \leq 0.92$ in all data unweighted, $0.87 \leq r \leq 0.92$ in all data weighted, and $0.85 \leq r \leq 0.90$ in match data unweighted.

Moreover, the results suggested that the adoption of IRT-based scores can contribute to quality control by removing potentially incorrect or inconsistent human scores, which leads to more consistent training of the neural networks. When the training set is either weighted or filtered based on IRT-generated scores, the agreement between human and automated scores was equal to or even improved compared to *all data unweighted* approach. While *all data unweighted* and IRT-based approaches showed equally high accuracy for SE71054 (r = 0.93) and ME72209 (r = 0.92), *match data unweighted* and *all data weighted* showed the highest level of accuracy for SE71077 (r = 0.96) and SE62005 (r = 0.87), respectively.

Table 4

Performance of Automated Scoring with ANNs

Item	All data unweighted	All data weighted	Match data unweighted
SE71054	0.93	0.92	0.93
SE71077	0.94	0.94	0.96
ME72209	0.92	0.92	0.90
SE62005	0.85	0.87	0.85
Average	0.91	0.91	0.91

* *Note*. Match: human score = IRT-based score; all data weighted: match = 2:1

For all four items, the confusion matrix for the approaches with the highest level of accuracy is presented below (see Tables 5-8). For the three items (SE71054, SE71077, and ME72209), false positive and false negative rates were commonly either equal to or less than 4%, while one difficult item (SE62005) showed a relatively high false positive rate (10%) and false negative rate (6%).

Table 5

Confusion Matrix for SE71054

			Human Score	
			0	1
Machine Score	0	All data unweighted	33%	3%
		Match data unweighted	34%	4%
	1	All data unweighted	4%	60%
		Match data unweighted	3%	59%

Table 6

Confusion Matrix for SE71077

			Human Score	
			0	1
Machine Score	0	Match data unweighted	39%	1%
	1	Match data unweighted	3%	57%

Table 7

Confusion Matrix for ME72209

			Human Score	
			0	1
Machine Score	0	All data unweighted	76%	4%
		All data weighted	76%	4%
	1	All data unweighted	4%	16%
		All data weighted	4%	16%

Table 8

Confusion Matrix for SE62005

			Human Score	
			0	1
Machine Score	0	All data unweighted	60%	6%
	1	All data unweighted	10%	24%

Discussion

This study has shown the feasibility of automated scoring for the CR items in ILSAs. Using four CR items from the TIMSS 2019 assessment, the study compared human scores with automated scores created from the ANN-based automated scoring model. There is substantial agreement between human and automated scoring for all four items. This suggests that automated scoring has the potential to support or substitute human scoring for short CR items. Remarkably, the adoption of IRT-based scores can be a promising strategy for improving the performance of automated scoring. When the ANN-based models were trained on weighted or filtered data based on IRT-generated scores, the classification accuracy increased for two items (SE71077 & SE62005). Although more items should be analyzed to generalize this finding in a future study, this implies that more improved performance could be achieved with the high-quality data which is weighted or filtered by IRT-based scores. It has been pointed out that achieving high-quality data is a vital step in supervised machine learning since errors in data can nullify the speed and accuracy of the performance (Breck et al., 2019; Prior et al., 2020; Riccio et al., 2020). Hence, the additional IRT-based scores introduced to weigh the data may prove useful for quality control.

Additionally, it should be noted that some misalignment of automated scores and human scores is inevitable as the classification accuracy was calculated based on human scores. Automated scores were compared against human scores, but some human raters generate incorrect or inconsistent scores. Therefore, the training based on single human ratings is less than ideal. In an ideal situation, only responses for which at least two human raters agree would be used in training. However, most testing programs apply double scoring only to a small fraction of all responses, mainly for estimating rater agreement. Also, the ANNs-based models depend on the bag-of-words model which only depicts the frequency of individual words in the data. Automated scoring determines the correctness of a student response using the feature matrix extracted from the bag-of-words model. This indicates that if a student writes a correct answer with only a few or no commonly used keywords, it can be possibly scored as incorrect. Further studies on addressing the inconsistency of human scoring will contribute to a more correct evaluation of automated scoring.

The advantage of ANN-based automated scoring is that it is expected to improve the accuracy and consistency of scoring while reducing the cost, time, and human efforts involved in training human raters. Despite such resource-intensive training, achieving high inter-rater reliability becomes more challenging when scoring large volumes of student responses in multilingual international assessments. Automated scoring can be generalized to multilingual responses with neural machine translation such as Google Translation API which supports over 100 languages. Translation of non-English language to English can be helpful to address potential problems associated with relatively small datasets of non-English language. Extensive quality control for translation is needed for quality assurance. Furthermore, automated scoring encourages students to review, revise, and improve their responses as this technology enables instant feedback (Wilson, 2017; Wang et al., 2021) while improving writing self-efficacy and performance (Wilson & Roscoe, 2020). This implies that automated scoring can be beneficial to writing instruction, beyond supporting or replacing human scoring.

One potential limitation of this study lies in the class imbalance of the two complex CR items (ME72209 and SE62005). They were highly skewed toward incorrect responses due to their complexity and difficulty. ME72209 became more imbalanced after cleaning the data based on the agreement between human and IRT-based scores. Although the data imbalance is common in a real-world context, it could lead to misclassification due to the bias towards a majority class (Feng et al., 2018; Hassib et al., 2019; Huang et al., 2018). Future research could tackle the issue of imbalance with various methods including data-level and algorithm-level strategies (Santos et al., 2018). Another limitation is that we did not provide a practical interpretation of student responses for which human and IRT-based scores disagreed on. In the next step of work, it will be worthwhile to score and examine those mismatched responses with a second human rater. Although the current study relied on a single human rater, a double human scoring would provide more reliable scores that can be used for training and comparisons.

Moreover, it should be noted that a few items showed slightly increased accuracy in the validation set than in the training set. For instance, SE71054 and ME72209 displayed higher accuracy in the validation set compared to the training set. This can probably be attributed to the spelling correction for which bad words in the validation set were replaced with good words from the training set. The spelling correction based on the training set may cause the overlap between the training and validation set and in turn, lead to slightly inflated performance

(Elangovan et al., 2021). Despite the unavoidable overlap, the benefit of this unique spelling correction is that bad words are more likely to be substituted with context-correct words. For instance, the word *squirrel* in SE71054 had 45 bad word variations in the data (e.g., *squal*, *squalrel*, *squrries*). Our spelling correction approach accurately corrected 80% of bad words, while the simple edit distance approach and *hunspell* were limited to 77.8% and 46.7%. In future research, a close analysis of different spelling correction methods would be a fruitful investigation.

Conclusion

Automated scoring is a feasible and practical alternative to human scoring while reducing many challenges required for training human raters. This study provides empirical evidence for the use of ANN-based automated scoring for short CR items in ILSAs. Not only did human and automated scores show very high agreement, but their agreement also increased more when ANNs were trained and tested on the data where human and IRT-expected scores matched. The next step will be to explore the scalability of this automated scoring to more CR items with varying difficulty and complexity as well as to multilingual student responses.

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

Agatonovic-Kustrin, S., & Beresford, R. (2000). Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. *Journal of Pharmaceutical and Biomedical Analysis*, *22*(5), 717-727. https://doi.org/10.1016/S0731-7085(99)00272-1

Alhawiti, K. M. (2014). Natural language processing and its use in education. *International Journal of Advanced Computer Science and Applications*, *5*. https://doi.org/10.14569/IJACSA.2014.051210

Attali, Y. (2007). Construct validity of e-rater® in scoring Toefl® essays. *ETS Research Report Series*, *2007*(1), i–22. https://doi.org/10.1002/j.2333-8504.2007.tb02063.x

Attali, Y., Powers, D., Freedman, M., Harrison, M., Obetz, S. (2008). Automated scoring of short-answer open-ended GRE subject test items. *ETS Research Report Series*, *2008*(1), i-22. https://doi.org/10.1002/j.2333-8504.2008.tb02106.x

Bahja, M. (2020). Natural language processing applications in business. *E-Business-Higher Education and Intelligence Applications*. IntechOpen. https://doi.org/10.5772/intechopen.92203

Bejar, I. I. (2012). Rater cognition: Implications for validity. *Educational Measurement: Issues and Practice*, *31*(3), 2-9. https://doi.org/10.1111/j.1745-3992.2012.00238.x

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., & Matsuo, A. (2018). quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software*, *3*(30), 774. https://doi.org/10.21105/joss.00774

Benoit, K., Watanabe, K., Wang, H., Lua, J. W., & Kuha, J. (2021). Package ‘quanteda. textstats’. *Research Bulletin*, 27(2), 37-54. https://cran.r-project.org/web/packages/quanteda.textstats/index.html

Boulis, C., & Ostendorf, M. (2005). Text classification by augmenting the bag-of-words representation with redundancy-compensated bigrams. *Proceedings of the international workshop in feature selection in data mining* (pp. 9-16). Citeseer. http://www.icsi.berkeley.edu/pubs/speech/bagofwords05.pdf

Braun, H. I., Bennett, R. E., Frye, D., & Soloway, E. (1990). Scoring constructed responses using expert systems. *Journal of Educational Measurement*, *27*(2), 93–108. https://doi.org/10.1111/j.1745-3984.1990.tb00736.x

Breck, E., Polyzotis, N., Roy, S., Whang, S. E., & Zinkevich, M. (2019). Data validation for machine learning. *Conference on systems and machine learning*. https://mlsys.org/Conferences/2019/doc/2019/167.pdf

Brown, J. D., & Hudson, T. (1998). The alternatives in language assessment. *TESOL Quarterly*, *32*(4), 653–675. https://doi.org/10.2307/3587999

Bryant, W. (2017). Developing a strategy for using technology-enhanced items in large-scale standardized tests. *Practical Assessment, Research, and Evaluation*, *22*(1), 1. https://doi.org/10.7275/70yb-dj34

Cardwell, R., LaFlair, G. T., & Settles, B. (2021). *Duolingo English test: Technical Manual*. Duolingo, Inc. https://englishtest.duolingo.com/research

Chary, M., Parikh, S., Manini, A. F., Boyer, E. W., & Radeos, M. (2019). A review of natural language processing in medical education. *Western Journal of Emergency Medicine*, *20*(1), 78. https://10.5811/westjem.2018.11.39725

Correnti, R., Matsumura, L. C., Wang, E., Litman, D., Rahimi, Z., & Kisa, Z. (2020). Automated scoring of students' use of text evidence in writing. *Reading Research Quarterly*, *55*(3), 493–520. https://doi.org/10.1002/rrq.281

Darling-Hammond, L. & Adamson, F. (2010). *Beyond basic skills: The role of performance assessment in achieving 21st century standards of learning*. Stanford, CA: Stanford University, Stanford Center for Opportunity Policy in Education. https://edpolicy.stanford.edu/library/publications/1462

Dikli, S. (2006). An overview of automated scoring of essays. *The Journal of Technology, Learning and Assessment*, *5*(1). https://ejournals.bc.edu/index.php/jtla/article/view/1640

Dongare, A. D., Kharde, R. R., & Kachare, A. D. (2012). Introduction to artificial neural network. *International Journal of Engineering and Innovative Technology (IJEIT)*, *2*(1), 189-194.

Elangovan, A., He, J., & Verspoor, K. (2021). Memorization vs. generalization: Quantifying data leakage in NLP performance evaluation. *arXiv preprint arXiv:2102.01818*.

Enright, M. K., & Quinlan, T. (2010). Complementing human judgment of essays written by English language learners with e-rater® scoring. *Language Testing*, *27*(3), 317–334. https://doi.org/10.1177/0265532210363144

Federer, M. R., Nehm, R. H., Opfer, J. E., & Pearl, D. (2015). Using a constructed-response instrument to explore the effects of item position and item features on the assessment of students' written scientific explanations. *Research in Science Education*, *45*(4), 527-553.. https://doi.org/10.1007/s11165-014-9435-9

Feng, W., Huang, W., & Ren, J. (2018). Class imbalance ensemble learning based on the margin theory. *Applied Sciences*, *8*(5), 815. https://doi.org/10.3390/app8050815

Fishbein, B., Foy, P., & Tyack, L. (2020). Reviewing the TIMSS 2019 achievement item statistics. In M. O. Martin, M. von Davier, & I. V. S. Mullis (Eds.), *Methods and Procedures: TIMSS 2019 Technical Report* (pp. 10.1-10.70). Retrieved from Boston College, TIMSS & PIRLS International Study Center website: https://timssandpirls.bc.edu/timss2019/methods/chapter-10.html

Foy, P., Fishbein, B., von Davier, M., & Yin, L. (2020). Implementing the TIMSS 2019 scaling methodology. In M. O. Martin, M. von Davier, & I. V. S. Mullis (Eds.), *Methods and Procedures: TIMSS 2019 Technical Report* (pp. 12.1–12.146). Boston College, TIMSS & PIRLS International Study Center. https://timssandpirls.bc.edu/ timss2019/methods/chapter-12.html

Flor, M., & Hao, J. (2021). Text mining and automated scoring. *Computational Psychometrics: New Methodologies for a New Generation of Digital Learning and Assessment* (pp. 245-262). Springer, Cham. https://doi.org/10.1007/978-3-030-74394-9_14

Ghag, K. V., & Shah, K. (2015). Comparative analysis of effect of stopwords removal on sentiment classification. *2015 International conference on computer, communication and control (IC4)* (pp. 1-6). https://doi.org/10.1109/IC4.2015.7375527

Gupta, N., Mujumdar, S., Patel, H., Masuda, S., Panwar, N., Bandyopadhyay, S., Mehta, S., Guttula, S., Afzal, S., Sharma Mittal, R., & Munigala, V. (2021). Data quality for machine learning tasks. *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, (pp. 4040–4041). https://doi.org/10.1145/3447548.3470817

Ha, M., & Nehm, R. H. (2016). The impact of misspelled words on automated computer scoring: A case study of scientific explanations. *Journal of Science Education and Technology*, *25*(3), 358-374. https://doi.org/10.1007/s10956-015-9598-9

Ha, M. (2016). Examining the validity of history-of-science-based evolution concept assessment and exploring conceptual progressions by contexts. *Journal of the Korean Association for Science Education*, *36*(3), 509-517. https://doi.org/10.14697/JKASE.2016.36.3.0509

Hancock, G. R. (1994). Cognitive complexity and the comparability of multiple-choice and constructed-response test formats. *The Journal of Experimental Education*, *62*(2), 143-157. https://doi.org/10.1080/00220973.1994.9943836

Harris, C. J., Krajcik, J. S., Pellegrino, J. W., & DeBarger, A. H. (2019). Designing knowledge‐in‐use assessments to promote deeper learning. *Educational Measurement: Issues and Practice*, *38*(2), 53-67. https://doi.org/10.1111/emip.12253

Hassib, E. M., El-Desouky, A. I., El-Kenawy, E. S. M., & El-Ghamrawy, S. M. (2019). An imbalanced big data mining framework for improving optimization algorithms performance. *IEEE Access*, *7*, 170774-170795. https://doi.org/10.1109/ACCESS.2019.2955983

Hearst, M. A. (2000). The debate on automated essay grading. *IEEE Intelligent Systems and Their Applications*, *15*(5), 22–37. https://doi.org/10.1109/5254.889104

Higgins, D., Xi, X., Zechner, K., & Williamson, D. (2011). A three-stage approach to the automated scoring of spontaneous spoken responses. *Computer Speech & Language*, *25*(2), 282–306. https://doi.org/10.1016/j.csl.2010.06.001

Huang, J. W., Chiang, C. W., & Chang, J. W. (2018). Email security level classification of imbalanced data using artificial neural network: The real case in a world-leading enterprise. *Engineering Applications of Artificial Intelligence*, *75*, 11–21. https://doi.org/10.1016/j.engappai.2018.07.010

Hussein, M. A., Hassan, H., & Nassef, M. (2019). Automated language essay scoring systems: A literature review. *PeerJ Computer Science*, *5*, e208. https://doi.org/10.7717/peerj-cs.208

Islam, M. M., & Hoque, A. L. (2010). Automated essay scoring using generalized latent semantic analysis. *2010 13th International conference on computer and information technology (ICCIT)* (pp. 358-363). IEEE. https://doi.org/10.1109/ICCITECHN.2010.5723884

Jivani, A. G. (2011). A comparative study of stemming algorithms. *International Journal of Computer Technology and Applications*, *2*(6), 1930-1938. https://www.semanticscholar.org/paper/A-Comparative-Study-of-Stemming-Algorithms-Jivani/4dbc8da1e4d23e9e7a9b966bc7ee547b2faac3e0

Jodoin, M. G. (2003). Measurement efficiency of innovative item formats in computer‐based testing. *Journal of Educational Measurement*, *40*(1), 1-15. https://doi.org/10.1111/j.1745-3984.2003.tb01093.x

Karpathy, A. (2015). The unreasonable effectiveness of recurrent neural networks. *Andrej Karpathy blog*. http://karpathy.github.io/2015/05/21/rnn-effectiveness/

Keevers, T. L. (2019). Cross-validation is insufficient for model validation. *Joint and Operations Analysis Division, Defence Science and Technology Group: Victoria, Australia*.

Kersting, N. B., Sherin, B. L., & Stigler, J. W. (2014). Automated scoring of teachers' open-ended responses to video prompts: Bringing the classroom-video-analysis assessment to scale. *Educational and Psychological Measurement*, *74*(6), 950-974. https://doi.org/10.1177/0013164414521634

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., ... & Team, R. C. (2020). *Package 'caret': Classification and regression training*. https://cran.r-project.org/web/packages/caret/caret.pdf

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211–240. https://doi.org/10.1037/0033-295X.104.2.211

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, *25*, 259–284. https://doi.org/10.1080/01638539809545028

LaVoie, N., Parker, J., Legree, P. J., Ardison, S., & Kilcullen, R. N. (2020). Using latent semantic analysis to score short answer constructed responses: Automated scoring of the consequences test. *Educational and Psychological Measurement*, *80*(2), 399-414. https://doi.org/10.1177/0013164419860575

Lee, H. S., Pallant, A., Pryputniewicz, S., Lord, T., Mulholland, M., & Liu, O. L. (2019). Automated text scoring and real‐time adjustable feedback: Supporting revision of scientific arguments involving uncertainty. *Science Education*, *103*(3), 590-622. https://doi.org/10.1002/sce.21504

Lissitz, R. W., Hou, X., & Slater, S. C. (2012). The contribution of constructed response items to large scale assessment: Measuring and understanding their impact. *Journal of Applied Testing Technology*, *13*(3). https://eric.ed.gov/?id=EJ1001221

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

Liu, O. L., Rios, J. A., Heilman, M., Gerard, L., & Linn, M. C. (2016). Validation of automated scoring of science assessments. *Journal of Research in Science Teaching*, *53*(2), 215–233. https://doi.org/10.1002/tea.21299

Liu, S., & Kunnan, A. J. (2016). Investigating the application of automated writing evaluation to Chinese undergraduate English majors: A case study of WriteToLearn. *Calico Journal*, *33*(1), 71–91. https://doi.org/10.1558/cj.v33i1.26380

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading: Addison-Wesley. https://psycnet.apa.org/record/1968-35040-000

Madnani, N., Burstein, J., Sabatini, J., & O’Reilly, T. (2013). Automated Scoring of Summary-Writing Tasks Designed to Measure Reading Comprehension. *Proceedings of the 8th workshop on innovative use of natural language processing for building educational applications* (pp. 163-168). Atlanta, GA: Association for Computational Linguistics. https://files.eric.ed.gov/fulltext/ED603960.pdf

Maestrales, S., Zhai, X., Touitou, I., Baker, Q., Schneider, B., & Krajcik, J. (2021). Using machine learning to score multi-dimensional assessments of chemistry and physics. *Journal of Science Education and Technology*, *30*(2), 239-254. https://doi.org/10.1007/s10956-020-09895-9

Martin, M. O., Mullis, I. V. S., & Hooper, M. (Eds.). (2017). Methods and Procedures in PIRLS 2016. Retrieved from Boston College, TIMSS & PIRLS International Study Center website: https://timssandpirls.bc.edu/publications/pirls/2016-methods.html

Martin, M. O., von Davier, M., & Mullis, I. V. S. (Eds.). (2020). *Methods and Procedures: TIMSS 2019 Technical Report*. Boston College, TIMSS & PIRLS International Study Center. https://timssandpirls.bc.edu/timss2019/methods

McClellan, C. A. (2010). Constructed-response scoring: Doing it right. *R&D Connections*, *13*. 1-7. https://www.ets.org/Media/Research/pdf/RD_Connections13.pdf

Miller, T. (2003). Essay assessment with latent semantic analysis. *Journal of Educational Computing Research*, *29*(4), 495-512. https://doi.org/10.2190/W5AR-DYPW-40KX-FL99

O'Leary, M., Scully, D., Karakolidis, A., & Pitsia, V. (2018). The state-of-the-art in digital technology-based assessment. *European Journal of Education*, *53*(2), 160–175. https://doi.org/10.1111/ejed.12271

Oliinyk, V. A., Vysotska, V., Burov, Y., Mykich, K., & Fernandes, V. B. (2020). Propaganda detection in text data based on NLP and machine learning. *MoMLeT+ DS* (pp. 132-144). http://ceur-ws.org/Vol-2631/paper10.pdf

Ooms, J. (2019). *The hunspell package: high-performance stemmer, tokenizer, and spell checker for R*. https://cran.r-project.org/web/packages/hunspell/vignettes/intro.html

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

Prior, F., Almeida, J., Kathiravelu, P., Kurc, T., Smith, K., Fitzgerald, T. J., & Saltz, J. (2020). Open access image repositories: High-quality data to enable machine learning research. *Clinical Radiology*, *75*(1), 7–12. https://doi.org/10.1016/j.crad.2019.04.002

Raschka, S. (2018). Model evaluation, model selection, and algorithm selection in machine learning. *arXiv preprint arXiv:1811.12808*.

Riccio, V., Jahangirova, G., Stocco, A., Humbatova, N., Weiss, M., & Tonella, P. (2020). Testing machine learning based systems: A systematic mapping. *Empirical Software Engineering*, *25*(6), 5193–5254. https://doi.org/10.1007/s10664-020-09881-0

Rokade, A., Patil, B., Rajani, S., Revandkar, S., & Shedge, R. (2018). Automated grading system using natural language processing. *2018 Second international conference on inventive communication and computational technologies (ICICCT)* (pp. 1123-1127). IEEE. https://doi.org/10.1109/ICICCT.2018.8473170.

Romanov, A., Lomotin, K., & Kozlova, E. (2019). Application of natural language processing algorithms to the task of automatic classification of Russian scientific texts. *Data Science Journal*, *18*(1). https://datascience.codata.org/articles/10.5334/dsj-2019-037/

Santos, M. S., Soares, J. P., Abreu, P. H., Araujo, H., & Santos, J. (2018). Cross-validation for imbalanced datasets: avoiding overoptimistic and overfitting approaches [research frontier]. *IEEE Computational Intelligence Magazine*, *13*(4), 59-76. https://doi.org/10.1109/MCI.2018.2866730

Sazli, M. H. (2006). A brief review of feed-forward neural networks. *Communications Faculty of Sciences University of Ankara Series A2-A3 Physical Sciences and Engineering*, *50*(1).

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

Shao, Y., Taylor, S., Marshall, N., Morioka, C., & Zeng-Treitler, Q. (2018). Clinical text classification with word embedding features vs. bag-of-words features. *2018 IEEE*

International conference on big data (Big Data) (pp. 2874-2878). IEEE. https://doi.org/10.1109/BigData.2018.8622345

Shermis, M. D., & Burstein, J. (Eds.). (2013). *Handbook of automated essay evaluation: Current applications and new directions.* Routledge/Taylor & Francis Group. https://psycnet.apa.org/record/2013-15323-000

Shermis, M. D., Burstein, J., Higgins, D., & Zechner, K. (2010). Automated essay scoring: Writing assessment and instruction. *International Encyclopedia of Education*, *4*(1), 20-26. https://doi.org/10.1016/B978-0-08-044894-7.00233-5

Sorin, V., Barash, Y., Konen, E., & Klang, E. (2020). Deep learning for natural language processing in radiology—fundamentals and a systematic review. *Journal of the American College of Radiology*, *17*(5), 639-648. https://doi.org/10.1016/j.jacr.2019.12.026

Sukkarieh, J., & Stoyanchev, S. (2009, August). Automating model building in c-rater. *Proceedings of the 2009 workshop on applied textual inference (TextInfer)* (pp. 61-69). https://aclanthology.org/W09-2509

Uysal, A. K., & Gunal, S. (2014). The impact of preprocessing on text classification. *Information Processing & Management*, *50*(1), 104-112. https://doi.org/10.1016/j.ipm.2013.08.006

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*. http://arxiv.org/abs/1706.03762 von Davier, M. (2018). Automated item generation with recurrent neural networks. *Psychometrika*, *83*(4), 847-857. https://doi.org/10.1007/s11336-018-9608-y

von Davier, M. (2019). Training optimus prime, MD: generating medical certification items by fine-tuning OpenAI's gpt2 transformer model. *arXiv preprint arXiv:1908.08594*.

von Davier, M. (2020). TIMSS 2019 scaling methodology: Item response theory, population models, and linking across modes. In Martin, M. O., von Davier, M., & Mullis, I. V. S. (Eds.), *Methods and Procedures: TIMSS 2019 Technical Report* (pp.11.1-11.25). Boston College, TIMSS & PIRLS International Study Center. https://timssandpirls.bc.edu/timss2019/methods/pdf/T19_MP_Ch11-scaling-methodology.pdf

von Davier, M., Tyack, L., & Khorramdel, L. (2022). Automated scoring of graphical open-ended responses using artificial neural networks. *arXiv preprint arXiv:2201.01783*.

Wahlen, A., Kuhn, C., Zlatkin-Troitschanskaia, O., Gold, C., Zesch, T., & Horbach, A. (2020). Automated scoring of teachers' pedagogical content knowledge–a comparison between human and machine scoring. *Frontiers in Education* (p. 149). Frontiers. https://doi.org/10.3389/feduc.2020.00149

Wang, C., Liu, X., Wang, L., Sun, Y., & Zhang, H. (2021). Automated scoring of Chinese grades 7–9 students' competence in interpreting and arguing from evidence. *Journal of Science Education and Technology*, *30*(2), 269-282. https://doi.org/10.1007/s10956-020-09859-z

Wang, J. (2020). Application of text clustering in automatic scoring of college English composition. *2020 2nd International conference on information technology and computer application (ITCA)* (pp. 598-603). IEEE. https://doi.org/10.1109/ITCA52113.2020.00131

Wang, S. C. (2003). Artificial neural network. *Interdisciplinary Computing in Java Programming* (pp. 81-100). Springer, Boston, MA. https://doi.org/10.1007/978-1-4615-0377-4_5

Wesolowski, M., & Suchacz, B. (2012). Artificial neural networks: theoretical background and pharmaceutical applications: A review. *Journal of AOAC International*, *95*(3), 652-668. https://doi.org/10.5740/jaoacint.SGE_Wesolowski_ANN

Williamson, D. M., Bejar, I. I., & Hone, A. S. (1999). 'Mental model' comparison of automated and human scoring. *Journal of Educational Measurement*, *36*(2), 158-184. https://doi.org/10.1111/j.1745-3984.1999.tb00552.x

Williamson, D. M., Bejar, I. I., & Sax, A. (2004). Automated tools for subject matter expert evaluation of automated scoring. *Applied Measurement in Education*, *17*(4), 323-357. https://doi.org/10.1207/s15324818ame1704_1

Wilson, J. (2017). Associated effects of automated essay evaluation software on growth in writing quality for students with and without disabilities. *Reading and Writing*, *30*(4), 691-718. https://doi.org/10.1007/s11145-016-9695-z

Wilson, J., & Roscoe, R. D. (2020). Automated writing evaluation and feedback: Multiple metrics of efficacy. *Journal of Educational Computing Research*, *58*(1), 87-125. https://doi.org/10.1177/0735633119830764

Yagis, E., De Herrera, A. G. S., & Citi, L. (2019, November). Generalization performance of deep learning models in neurodegenerative disease classification. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (pp. 1692-1698). IEEE.

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

CHAPTER 3. PAPER 2

Jung, J.Y., Tyack, L. & von Davier, M. Combining machine translation and automated scoring in international large-scale assessments. Large-scale Assessments in Education, *12*, 10 (2024). https://doi.org/10.1186/s40536-024-00199-7

Combining Machine Translation and Automated Scoring in International Large-scale Assessments

Abstract

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. 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. 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. This study highlights that automated scoring

integrated with the recent machine translation holds great promise for consistent and resource-efficient scoring in ILSAs.

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; Wolfe & McVay, 2012; von Davier et al., 2023).

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 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 natural language processing (NLP) tools and research are predominantly focused on English (Hovy et al., 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 (LRLs) (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).

Moreover, the abundance of responses collected in ILSAs has historically posed a challenge for scoring CR items. Modern NLP and 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 & Aditya et al., 2016; Prasanna & Rao, 2018; Sutskever et al., 2011; Wang et al., 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 written response 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.

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

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.

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.

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

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 free 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 and selection of machine translation 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.

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).

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)

$$Exact\ Match\ Ratio\ (classification\ agreement)\ =\ \frac{BI+BC}{BI+BC+D1+D2}$$

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:

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

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.

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

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.

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

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 6). 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 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

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.

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

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).

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

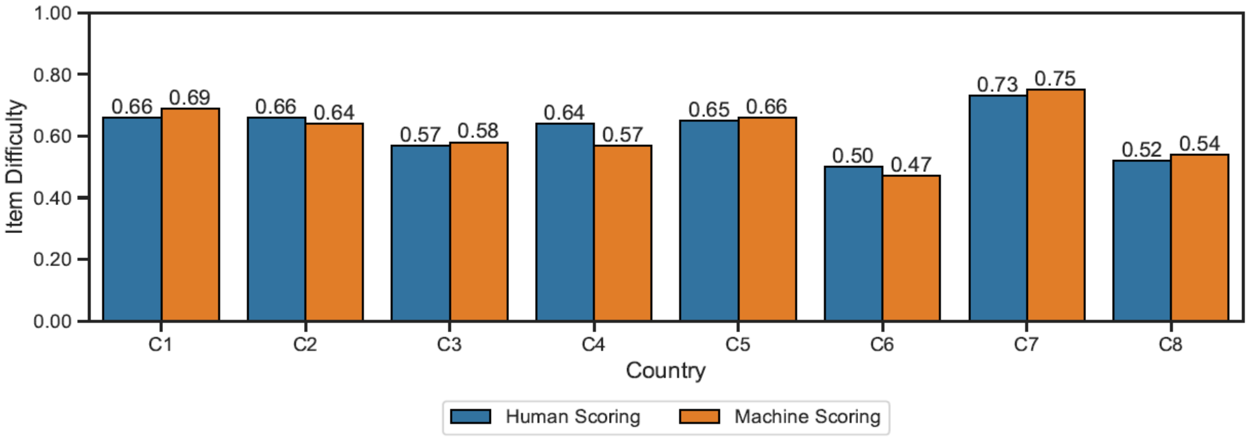
Note. AITC = Adjusted Item-total Correlation

The AITC generally displayed consistent patterns across countries and scoring methods, with a slightly stronger correlation in human scoring (see Table 12 attached as an additional file). 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).

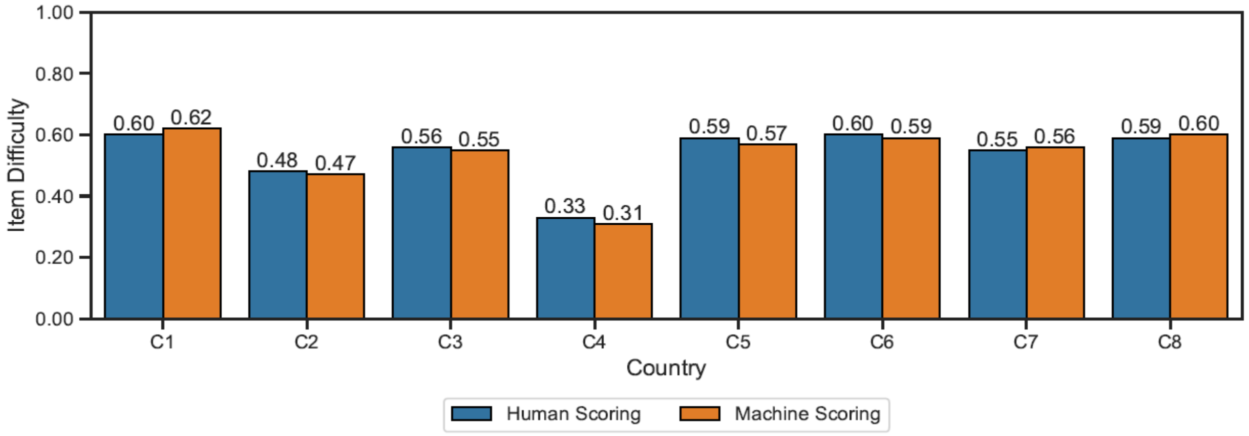
Next, the overall patterns of country-by-item difficulty remained consistent across the scoring methods (see Figures 1-6). Importantly, even uncommon patterns were maintained across the scoring method (refer to Figure 3). In human scoring, Item 3 was relatively easy for students in C7 (r_{human} = 0.54) 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.

Figure 1 *County-by-Item Difficulty of Item 1*



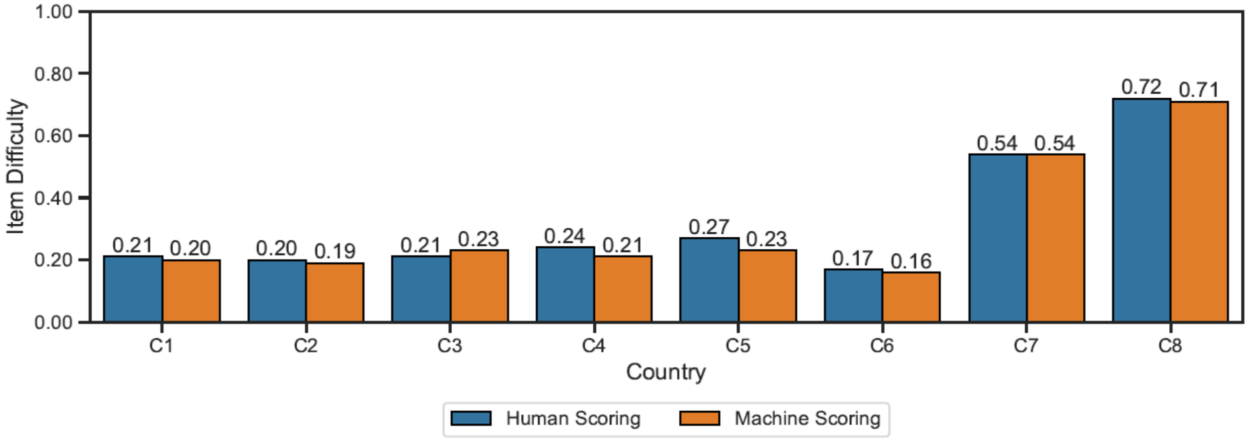
Note. C1 & C2 = German-speaking countries; C3 = French-speaking country; C4 = Turkish-speaking country; C5 = English-speaking country; C6 & C7 = Chinese-speaking countries; C8 = Korean-speaking country

Figure 2 *Country-by-Item Difficulty of Item 2*



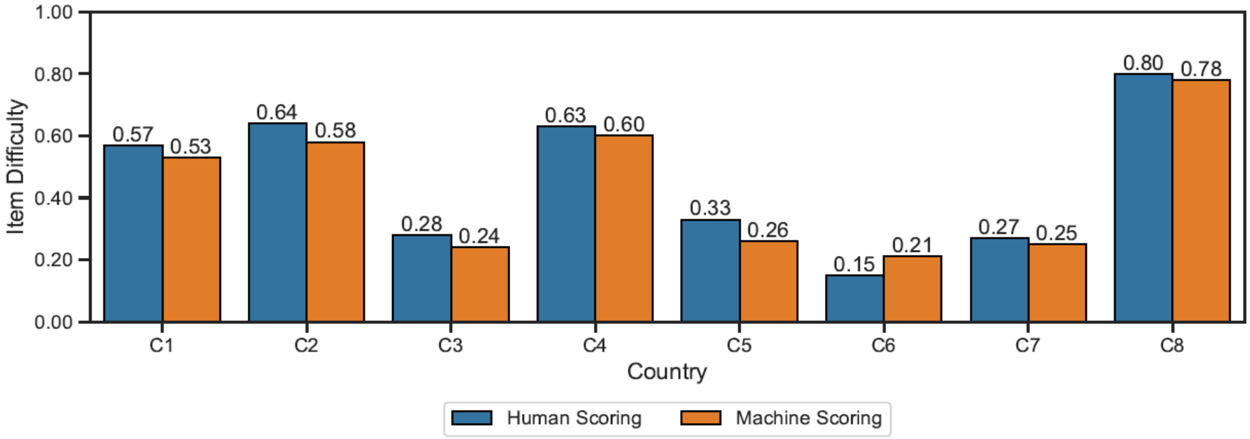
Note. C1 & C2 = German-speaking countries; C3 = French-speaking country; C4 = Turkish-speaking country; C5 = English-speaking country; C6 & C7 = Chinese-speaking countries; C8 = Korean-speaking country

Figure 3 *Country-by-Item Difficulty of Item 3*



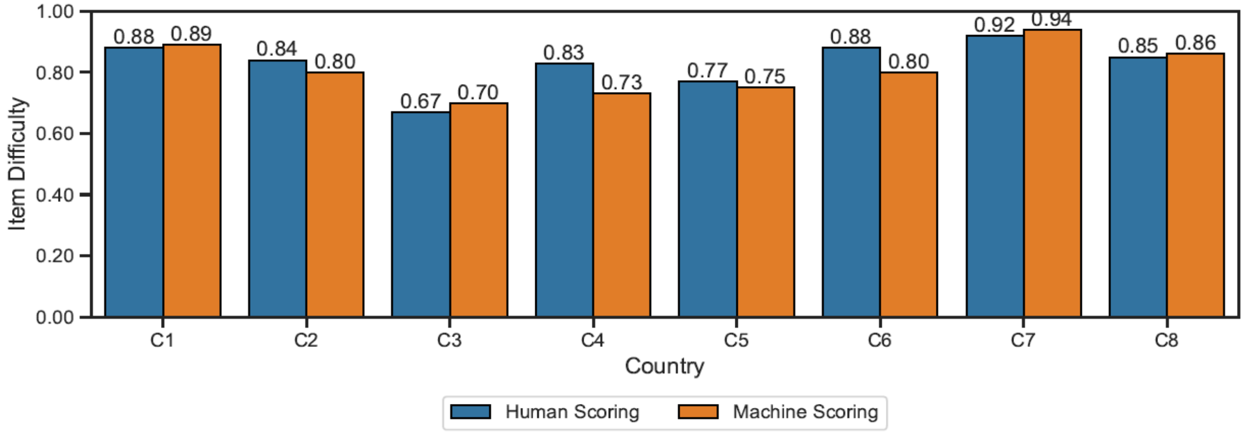
Note. C1 & C2 = German-speaking countries; C3 = French-speaking country; C4 = Turkish-speaking country; C5 = English-speaking country; C6 & C7 = Chinese-speaking countries; C8 = Korean-speaking country

Figure 4 *Country-by-Item Difficulty of Item 4*



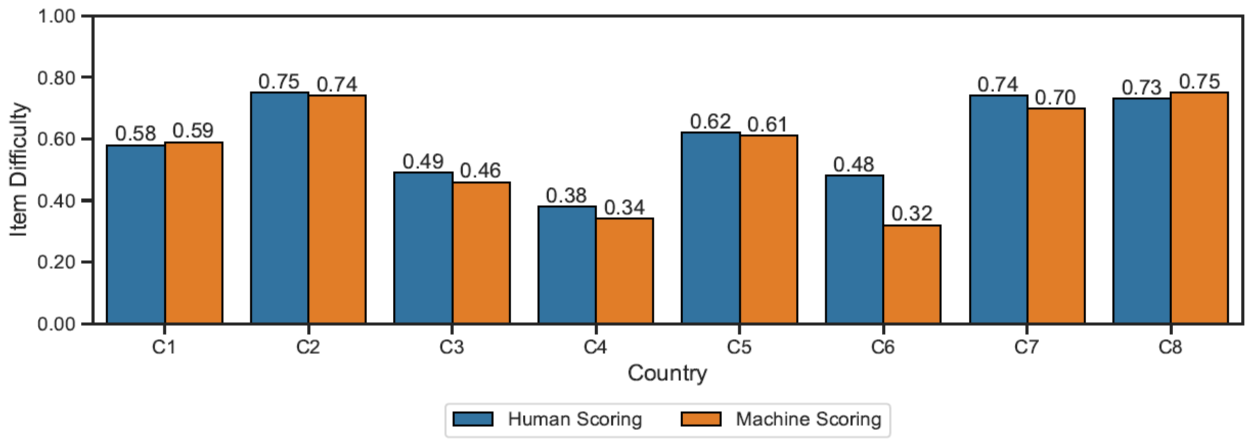
Note. C1 & C2 = German-speaking countries; C3 = French-speaking country; C4 = Turkish-speaking country; C5 = English-speaking country; C6 & C7 = Chinese-speaking countries; C8 = Korean-speaking country

Figure 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

Figure 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

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 seconds by Google Translate and ChatGPT, respectively. Running the ANNs per item took approximately 7.50 minutes 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 (Mikolov et al., 2013; Chen et al., 2019).

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 (κ = 0.97 – 1.00) in this study, we observed slight within-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.

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 Study

We observed that the differences between human and machine scores were derived from various factors, not just the error of FNN 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.

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). New York, NY: 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, M. T., & 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://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: Volume 1, Long Papers* (pp. 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 (2023a). Models. *OpenAI*. https://platform.openai.com/docs/models/overview

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

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

CHAPTER 3. PAPER 3

Towards the Implementation of Automated Scoring in International Large-scale Assessments: Scalability and Quality Control

Abstract

Automated scoring has received considerable attention in educational measurement, even before the age of artificial intelligence. However, its application to constructed response (CR) items in international large-scale assessments (ILSAs) has remained a challenge, primarily due to the difficulty of handling multilingual responses spanning many languages. This study addresses this challenge by investigating two machine learning approaches — supervised and unsupervised learning — for scoring multilingual responses. We explored various scoring methods to assess three science CR items from TIMSS 2023 across all participating countries and 42 languages. The results showed that the supervised learning approach, particularly combining multiple machine translations with artificial neural networks (MMT_ANNs), showed comparable performance to human scoring. The MMT_ANN model demonstrated impressive accuracy, correctly classifying up to 94.88% of responses across all languages and countries. This remarkable performance can be attributed to MMT_ANNs providing more suitable translations at both individual response and language levels. Furthermore, MMT_ANNs consistently generated accurate scores for identical or borderline responses within and across countries. These findings indicate the potential of automated scoring as an accurate and cost-effective measure for quality control in ILSAs, reducing the need to hire more human raters to ensure scoring reliability.

1. Introduction

Constructed response (CR) items are considered crucial in assessing students' thinking and reasoning abilities (Bennett, 1991; Liu et al., 2011; Liu et al., 2014; Mao et al., 2018; McClellan, 2010). However, their use in international large-scale assessments (ILSAs) is often limited by the subjective and costly nature of human scoring (Braun et al., 1990). Consequently, multiple-choice (MC) items have become the predominant choice in ILSAs due to their reliability and amenability to rapid scoring (Livingston, 2009). In contrast, the human scoring of CR items includes constructing comprehensive scoring guides, recruiting and training human raters, and monitoring the scoring process to achieve high scoring reliability (Bennett, 1991; Martin et al., 2020). Despite extensive training and monitoring, scoring inconsistencies and subjectivity persist due to rater effects, including severity, leniency, fatigue, and interpretation differences (Bejar, 2012; Myford & Wolfe, 2003; Wang & von Davier, 2014; Zhang, 2013). Moreover, the diverse language responses in ILSAs, which span over 60 countries, pose significant challenges in maintaining scoring standards across different languages and regions (Okubo et al., 2023). Automated scoring can be used to address these challenges and promote the broader use of CR items in ILSAs while reducing the time and cost associated with hiring additional human raters.

There have been many efforts to implement the automated scoring of CR items across various assessments (Gilson et al., 2023; LaVoie et al., 2020; Lottridge et al., 2018; Liu et al., 2014; Shermis & Wilson, 2024). However, most of these endeavors have focused on monolingual contexts, often centering on English responses. The rapid advancements in artificial intelligence (AI), including artificial neural networks (ANNs), machine translation (MT), and large language models (LLMs), suggest the potential for automated scoring to extend to a multitude of languages and countries. Unlike earlier machine-supported scoring methods reliant on manual feature selection, ANNs can automatically detect patterns within data and be used for classification and prediction tasks (El Naqa & Murphy, 2015; Kim, 2014; Surden, 2021). Moreover, state-of-the-art MT engines can swiftly and accurately translate numerous languages. For instance, Google Translate and Meta's AI model support over 100 languages (Caswell, 2022; META, 2022), including low-resource languages (LRLs) which lack quality training corpus. Furthermore, LLMs such as OpenAI's ChatGPT and Google's Gemini can be used for text generation, prediction, translation, and summarization tasks. Jiao et al. (2023) found that ChatGPT could perform competitively with commercial translation engines such as Google Translate. These advancements in AI technology can expand automated scoring across many languages and countries while supporting human scoring validation.

This study explored multiple automated scoring methods using three science items from the 2023 cycle of the Trends in International Mathematics and Science Study (TIMSS). We aimed to demonstrate that automated scoring of CR items can be consistently and accurately implemented across all participating countries in ILSAs. The results can be valuable evidence supporting the scalability of automated scoring in numerous countries. Also, we propose automated scoring as an extra quality control measure to ensure reliable scoring of CR items in ILSAs. We addressed the following research questions:

1. Does automated scoring exhibit consistent and accurate performance across all participating countries and languages in TIMSS 2023?
2. How can MT tools be effectively used to expand the scope of automated scoring?

3. What causes discrepancies between human and automated scoring in a multilingual context?

2. Background

As assessments have transitioned to computer-based testing, which incorporates a variety of technology-enhanced items, implementing automated scoring in ILSAs has become increasingly important. TIMSS 2023, for example, shifted the assessment to a digital environment for all participating countries and included innovative items such as the Problem Solving and Inquiry (PSI) tasks. PSI items were designed to measure students' higher-order mathematics, science, and critical thinking skills in real-world scenarios (Mullis et al., 2021). Similarly, the 2015 cycle of the Programme for International Student Assessment (PISA) adopted a computer-based platform as its primary mode of assessment and featured many interactive CR items (OECD, 2024). The increased use of CR items is encouraging, given their importance in assessing students' reasoning and critical thinking through inquiry (Liu et al., 2011; McCarthy, 2005); CR items are believed to provide a more comprehensive understanding of student knowledge and thinking processes than MC items (Federer et al., 2015).

However, scoring a substantial volume of multilingual responses in ILSAs remains challenging. Human scoring CR items demands significant time, effort, and labor and introduces subjectivity and inconsistency (Bejar, 2012). While ILSAs assess human scoring reliability within and across countries, it is not feasible to assess all CR items. For instance, TIMSS is restricted to double-scoring 200 responses per item in each country and producing inter-rater reliability estimates. This approach is limited, as only a subset of responses is scored twice by independent human raters, potentially excluding challenging-to-score responses. Additionally, estimates may not reflect consistency across countries because human raters are trained individually by their country's representatives and might assign differing scores compared to raters from other countries (Martin et al., 2020). Automated scoring holds promise as an efficient method for producing additional scores for all student responses in ILSAs because it can apply the same scoring rules across countries while costing less and running faster than human raters.

Over the past five decades, numerous automated scoring engines for CR items have been developed with the progress in natural language processing (NLP) methodologies. In the United States, states like West Virginia, Louisiana, and Utah employ automated scoring to score tens of millions of student responses (Johnson & McCaffrey, 2023). Generally, states have found more advantages than drawbacks in automated scoring, though the transition has not always been seamless due to concerns around scoring errors. (Shermis & Lottridge, 2019). Moreover, testing companies have developed scoring engines for the operational scoring of millions of short or essay-length responses; some examples are ACT’s CRASE, Cambium’s Autoscore, the Educational Testing Service (ETS)’s e-rater, the Duolingo English Test, and Pearson’s Intelligent Essay Assessor. Specific to science items, the ETS introduced the c-rater and c-rater-ML for scoring short written responses. Research suggests that both engines perform comparably to human scoring (Liu et al., 2016; Sukkarieh & Blackmore, 2009).

Despite the increasing use of automated scoring in assessments, its implementation is often confined to monolingual contexts. Most NLP research focuses on high-resource languages like English, which feature in over 60% of NLP papers published (Cieri et al., 2016; Mielke et al., 2019). LRLs such as Albanian, Kazakh, and Hindi receive less attention due to the scarcity of large, high-quality corpora. Given that ILSAs can include several dozens of languages (and often more than 100 language versions), the shift to automated scoring in a multilingual context is necessary. However, applying automated scoring to written responses in ILSAs still needs to be explored, with only recent literature incorporating multiple languages. Shin et al. (2024) employed automated scoring to reading items using the Progress in International Reading Literacy Study (PIRLS) 2016 data across 14 languages. Likewise, Jung et al. (2024) applied automated scoring to six TIMSS 2019 items using responses from six different languages. While both studies found that automated scoring could be applied across different languages effectively in ILSAs, the languages examined were only a fraction of those used in the PIRLS and TIMSS assessments. Therefore, this study used data from all TIMSS languages to advance the multilingual automated scoring research and further evidence the applicability of this validation approach across all participating countries in ILSAs.

Recent advances in MT have significantly enhanced the feasibility of automated scoring for multilingual responses. The performance of MT significantly improved with the introduction of attention mechanisms in language models, which prioritize and focus on important inputs, mimicking the function of the human brain (Choi et al., 2018; Luong et al., 2015; Tu et al., 2016). Modern MT engines offer swift, accurate, and cost-effective translation, suggesting that translating non-English language responses into English and then proceeding with the subsequent analysis (i.e., automated scoring) in English can be a competitive strategy (Araújo et al., 2020; Arun et al., 2020; Horbach et al., 2023). The current study explored the promise of combining automated scoring with MT to score all multilingual responses from a selection of TIMSS 2023 science items.

3. Methods

3.1 Dataset

This study used three CR science items from TIMSS 2023. Items 1 and 2 assessed Grade 4 knowledge of Physical Science and Life Science, respectively, while Item 3 assessed Grade 8 knowledge of Physics. The dataset encompassed all participating countries in TIMSS 2023, totaling 52 countries for Items 1 and 2 and 40 countries for Item 3. Table 1 illustrates the distribution of 42 distinct languages in the dataset, with the number of language versions next to each entry. In TIMSS, each country team manages the translation of their unique materials. Thus, items administered in the same non-English language may have translation differences across countries. Each of the three items was scored dichotomously, with a score of 1 assigned to correct responses and a score of 0 to incorrect responses. The names of countries have been anonymized to respect privacy.

Table 1

Languages included in the TIMSS 2023 dataset

Language	# Versions	Language	# Versions	Language	# Versions
Albanian	3	Galician	1	Montenegrin	1
Arabic	9	Georgian	1	Nynorsk	1
Armenian	1	German	2	Polish	2
Azerbaijani	1	Greek	1	Portuguese	3

Basque	1	Hebrew	1	Romanian	1
Bokmal	1	Hungarian	2	Russian	5
Bosnian	1	Irish	1	Serbian	2
Catalan	1	Italian	1	Slovak	1
Chinese	3	Japanese	1	Slovene	1
Croatian	1	Karakalpak	1	Spanish	2
Czech	1	Kazakh	1	Swedish	2
Danish	1	Korean	1	Turkish	1
Dutch	2	Latvian	1	Uzbek	1
English	17	Lithuanian	1	Valencian	1
Finnish	1	Macedonian	1		
French	4	Malay	1		

Aberrant responses consisting solely of punctuations, numbers, single characters, or repetitive characters were considered meaningless and automatically received a machine score of 0. These responses were excluded from the automated scoring training and validation process. The remaining dataset was divided into training and test sets using an 80%:20% ratio across all countries. Table 2 presents the number of student responses automatically scored as incorrect (auto-scored) and those assigned to the training and test sets.

Table 2

Sample sizes for the auto-scored, training, and test sets

	Auto-score	Training	Test	Total
Item 1	314	32,166	8,038	40,518
Item 2	300	35,247	8,811	44,358
Item 3	688	25,556	6,390	32,634
Average	434	30,990	7,746	39,170

3.2 Supervised Learning

Supervised learning uses previously human-labeled data to train ANNs. We investigated three distinct combinations of MT and ANN automated scoring using Google Translate only, translation with ChatGPT (i.e., gpt-turbo-0125) only, and a mix of Google Translate and ChatGPT; in the following, we refer to these three applications as Google_ANNs, ChatGPT_ANNs, and MMT_ANNs. These three approaches will be explained in more detail below. Following MT, translated and native English responses were compiled, and training and testing were performed using ANNs via the Python *scikit-learn* package.

3.2.1 Machine Translation

Google Translate. Google Translate automatically identifies the original language of the responses and translates them into English. However, there were instances where Google Translate provided the original language response instead of an English translation. Also, it sometimes provided meaningless transliteration, the process of phonetically converting the words of a language into a foreign script (Prabhakar & Pal, 2018); for instance, the Arabic response "الجظيبي" was transliterated into "Al-Jazibi" or the misspelled Chinese response "地新影力" into "Dixin influence". These responses were deemed untranslated and removed from the training and test sets.

ChatGPT Translation. ChatGPT, or more generally, LLMs, can be used for machine translation as many of these models are trained on multilingual corpora (Jiao et al., 2023). To explore LLM capabilities, we utilized OpenAI's ChatGPT for MT. Because of potential privacy concerns, we did not use any identifiable information with ChatGPT. The data was fed to the system anonymously without any context that could identify a respondent. Before translation with ChatGPT, we standardized the format of responses by replacing multiple blank spaces or tabs with a single space. We then instructed ChatGPT to translate non-English responses based on a step-by-step prompt. We offered the context (item stem), an example of the required format in JSON, and an example of the output. For instance, if the original response was "물이 필요하다", ChatGPT would generate an output like {"물이 필요하다": "*It needs water*"}.

For untranslatable responses, ChatGPT was instructed to output "uncertain", as in the following example: {"뮤르필하다": "uncertain"}. This prevents ChatGPT from generating inaccurate translations or hallucinations for unclear or unrecognizable responses. Responses labeled "uncertain" underwent a re-translation process with a prompt specifically designed to translate potentially misspelled responses. Responses that remained "uncertain" after re-translation were categorized as non-translated and removed from the training and test sets.

Multiple Machine Translation. Our MMT_ANN approach combined translations from Google Translate and ChatGPT. We determined the final MT engine for each language group by comparing human-machine score agreements computed from Google-translated and ChatGPT-translated responses. Selecting the translation tool based on the training and human scorer agreement does not involve bleeding information from the testing data into the selection. Also, since TIMSS does not use machine scoring operationally for written response items, human scores are available as a legitimate training target. The machine-human agreement is maximized in the way that quality control measures can be assured based on the best possible MT choice and prediction model.. Finally, quality control measures could be taken if we found language samples that showed low agreement even after this best possible training to maximize the match. Following the best possible training, which maximizes the match between human and machine scores, we asked ourselves several questions to better understand what could be learned from country/language samples where agreement remained lower than expected.These questions were:

1. Is there an issue with MT for that language?
2. Is there an issue with the Bag-of-Word (BOW) being very different for that set of translated responses?
3. Is there an issue with a particular human scorer?
4. Is there a more general issue with the scoring guide in that country/language sample?

The proposed method allows answering these questions, which are more specific versions of the research questions provided above, focusing on the reasons behind the lower-than-expected human-machine agreement. For future operational use, other criteria could be used to select the tool for MMT_ANN, such as the number of untranslated responses and the similarity of the resulting vocabulary to the BOW of other languages, and untranslated responses.

While Google Translate served as the default MT method for all languages, ChatGPT was used instead for languages where ChatGPT-translated data yielded 2% or higher agreement with human scoring than Google-translated data. Once each language's final MT tool was selected, we adjusted translations at the individual response level. Alternative MT could be employed for individual responses if the final MT provided inappropriate translations. To illustrate the process, assume that Google Translate produced a translation identical to the original response or a translation where more than half of the words were in a non-English language (e.g., transliteration), then the translation was substituted with a ChatGPT translation. Conversely, if ChatGPT produced an "uncertain" label (could not translate the response) or meaningless output instead of an English translation, the translation was replaced with a Google translation. Responses that could not be translated by either method were filtered from the training and test sets and tallied to assess the overall success rate of translations.

3.2.2 Pre-processing, Bag-of-Words & ANNs

The joint training set was generated by combining translated and native English responses. Identical pre-processing steps, BOW, and ANNs were applied to three sets of multilingual data obtained from Google Translate (Google_ANNs), ChatGPT

(ChatGPT_ANNs), and MMT (MMT_ANNs). A set of pre-processing steps, including tokenization, lower-casing, spelling correction, and stemming, was applied. Misspelled words in the test set were corrected using a spelling correction dictionary extracted from the training set (Jung et al., 2022). Additionally, native English responses underwent an extra spelling correction using the *pyspellchecker* package in Python. This step was necessary because English responses contained more misspellings than translated non-English responses, which benefited from implicit spelling corrections applied through MT. This step cannot be controlled by comparing translations; different MT algorithms apply different levels of 'tolerating' or 'correcting' for misspellings so that even heavily misspelled terms may receive a translation that retains the intended meaning.

After pre-processing, a BOW key feature matrix was created from the training set, retaining only words appearing in at least 0.05% of the responses. BOW representation was used to represent all translated and English responses using a common key feature matrix. For example, the sentence "*It is wet*" could be transformed into {"*it*", "*is*", "*wet*"}. Given that the selected CR items typically require brief responses containing a few words, including a keyword, using the BOW can be considered an effective method for scoring these types of short responses. Next, fully connected feed-forward neural networks (FNNs), a type of ANN, were implemented using the *scikit-learn* package in Python. FNNs were constructed with three layers: the input layer, the hidden layer, and the output layer (Sazli, 2006). The neurons in this model were interconnected only in the direction from input to output, without cyclic connections between the layers. The BOW feature matrix was fed into the input layer and passed on to the output layer, which generates a machine score of 0 for incorrect responses and 1 for correct responses for test data. The trained FNNs were then applied to the independent test set for evaluation.

3.3 Unsupervised Learning

Traditionally, unsupervised learning (UL) refers to leveraging machine learning algorithms to identify latent structures or patterns within unlabeled data; it does not require human-scored data for training. In our study, we used LLMs for the same purpose, labeling responses as correct or incorrect without a one-to-one match with human responses. To produce machine scores, we provided the LLMs with the item stems and scoring guides, which also included examples of correct or incorrect student responses. This context was included to allow LLMs to provide reliable classifications for the student responses. We conducted our UL analysis using a two-step process: first, responses were translated and scored by ChatGPT, then majority scoring with fuzzy matching was applied to compensate for inconsistencies in ChatGPT performance.

3.3.1 ChatGPT Translation & Scoring

Before translation, we standardized the format of responses because ChatGPT is sensitive to input variations or repeated attempts at the same inquiry (Azaria et al., 2024). A series of preprocessing steps were implemented: punctuations such as forward slashes or brackets were removed, redundant blank spaces or tabs were replaced with a single space, and numbered lists of responses were converted into sentences. We set ChatGPT's temperature to 0.2, where a lower temperature closer to 0 indicates a more focused task and reduces creativity. Following these steps, we formatted translations and scores in JSON to ensure consistent output. Example outputs (e.g., *{"translated response": "Score: 1 or 0"}*) and context (the item stem) were also incorporated into the prompt. Scoring guides containing the criteria for awarding credit and example student responses were also included.

3.3.2 Majority Scoring under Fuzzy Matching

Students often use lexically similar but inexact keywords in their responses. This led to high frequencies of responses that were similar but not identical. This was particularly true for CR items requiring only a brief phrase or expecting a specific keyword as the correct answer. Several CR items in TIMSS are like this, asking for a specific scientific concept or keyword that shows student understanding. For example, the German word "*schwerkraft*", which means gravity, can be misspelled in several ways, including "*schwergraft*", "*schverkraft*", and "*schverkraft*". Similarly, for Danish, more than 30 misspelled variants of "*tyngdekraft*" (gravity) were observed, including "*tøndekraften*", "*tyntekraften*" and "*tungdekraften*". ChatGPT occasionally provided inconsistent translations or scores for these lexically similar responses. Overall, this was a rare occasion and mainly affected misspelled words, where some misspelled variants would be incorrectly translated based on the given question. To ensure scoring consistency, we decided to group lexically similar responses in the original language and assign them the same score regardless of translation.

Fuzzy matching was used to detect lexically similar (but not exactly matching) responses (Cayrol et al., 1982; Dubois & Prade, 1993). Fuzzy matching requires a prior criterion for the length of the text span to be matched. One criterion is edit distance, which is a metric representing the number of operations needed to transform one word into another, including replacements, insertions, or deletions (Levenshtein, 1966). In this study, responses within an edit distance of two from each of the top ten most frequently appearing unique responses were considered fuzzy-matched. For instance, misspelled variants like "*schwergraft*" and "*schverkraft*" would be grouped due to their lexical similarity to "*schwerkraft*," the most frequent word that represents a correct response in the German dataset for Item 1. Similarly, misspelled Danish words like "*tøndekraften,*" "*tyntekraften*" and "*tungdekraften*" were regarded as equivalent to "*tyngdekraft".*

Despite initially scoring all responses with ChatGPT, we adjusted the final machine score to ensure similar responses were scored consistently. Identical machine scores were given to the fuzzy-matched responses based on the majority scoring, which uses the "winning" class as the final score (Salminen et al., 2021). If more than half of the fuzzy-matched responses received a machine score of 1, then all corresponding responses were assigned a final machine score of 1(and vice versa). For example, about 85% of the fuzzy-matched responses for "*schwerkraft*" (e.g., "*schwergraft*" and "*schverkraft*") received a machine score of 1, while the remaining 15% (e.g., "*schertraft*" and "*schwer kraft*") received a machine score of 0. In this case, the final machine score for all fuzzy-matched responses (e.g., "*schwergraft,*" "*schverkraft,*" "*schertraft*" and "*schwer kraft*") was determined to be 1 based upon thethe majority score of 1 given to responses of "*schwerkraft".* This approach ensured consistent scoring for lexically similar but misspelled responses, overcoming the challenge of incorrect translation and improper scoring caused by spelling errors.

3.4 Evaluation Metrics

The performance of automated scoring was evaluated employing various metrics: MT performance, mean scores, standardized mean score difference (SMD), exact agreement (EA), and Cohen's kappa.

3.4.1 MT Performance: Untranslatable Responses

The percentage of untranslated responses across various MT methods were examined. Several causes could lead to untranslated responses in MT, including spelling errors, abbreviations, dialects/idiolects, and limited parallel data for LRLs. Given that our CR items required concise responses, the percentage of untranslated responses was considered a direct metric to evaluate MT performance.

3.4.2 Distributional differences: Mean & SMD

Mean scores and SMD were calculated for individual countries per item. SMD represents the mean difference between human and machine scores divided by the pooled standard deviation. An SMD value of 0 indicates no difference between human and machine scores. Positive values indicate automated scoring yields a higher mean score, while negative values suggest the opposite. The absolute value of SMD should not exceed 0.15 (Williamson, 2012). SMD is computed as:

$$SMD = \frac{\overline{X_M} - \overline{X_H}}{\sqrt{\frac{SD_M^2 + SD_H^2}{2}}}$$

Where $\overline{X_M}$ is the machine mean score, $\overline{X_H}$ is the human mean score, SD_M^2 is the variance of the machine score, and SD_H^2 is the variance of the human score.

3.4.3 Agreement Statistics: Exact Agreement (EA) & Cohen's Kappa

EA is a widely used metric that evaluates the percentage of exact agreement between two independent scores, as indicated by Both 0 (B0) and Both 1 (B1) in Table 3. We assessed EA values for both human-human scores and human-machine scores. Auto-scored responses were also considered when calculating final EA values.

Table 3

Confusion matrix in automated scoring

	Human Score 0	Human Score 1
Machine Score 0	Both 0 (B0)	Disagree (D1)
Machine Score 1	Disagree (D2)	Both 1 (B1)

$$EA = \frac{B0+B1}{B0+B1+D1+D2}$$

Cohen's kappa (κ) is a more stringent measure than EA as it assesses agreements beyond chance (Cohen, 1960). κ values range between -1 and 1, where 0 indicates agreement merely by chance. We adhered to the standards proposed by Landis and Koch (1977): values ≤ 0.00 are interpreted 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.

$$\kappa = \frac{P_o - P_c}{1 - P_c}$$

where P_o is the percentage of observed agreement, while P_c is the percentage of agreement expected by chance.

4. Results

4.1 Human Scoring Reliability

Two independent human raters scored approximately 200 responses for each item per country in TIMSS 2023 to evaluate the consistency of human scoring for the CR items; the exact agreement (EA) between the two human scores was calculated. Overall, human scoring exhibited high reliability across all three items, with average EAs ranging between 97.50% and 98.51%. Countries C03 on Item 2 (76.40%) and C22 on Item 3 (83.92%) demonstrated relatively moderate EA.

4.2 Evaluation of Automated Scoring Methods

4.2.1 MT Performance: Untranslated Responses

Table 4 displays the number and percentage of untranslated responses across various scoring methods. On average, MMT_ANNs exhibited the lowest rate of untranslated responses (0.81%), followed by ChatGPT_ANNs (1.39%), UL (1.49%), and Google_ANNs (2.64%). MMT_ANNs could translate almost all non-English responses by combining Google Translate and ChatGPT translations. Also, ChatGPT_ANNs and UL translated more responses than Google_ANNs, possibly due to ChatGPT's context-based translation approach.

Table 4

Number and percentage of untranslated responses across automated scoring methods

	Google_ANNs		ChatGPT_ANNs		MMT_ANNs		Unsupervised Learning	
	n	%	*n*	%	*n*	%	*n*	%
Item 1	1470	3.66%	464	1.15%	318	0.79%	792	1.97%
Item 2	767	1.74%	382	0.87%	223	0.51%	676	1.53%
Item 3	806	2.52%	690	2.16%	360	1.13%	308	0.96%
Average	1014	2.64%	512	1.39%	300	0.81%	592	1.49%

4.2.2 Distributional Differences: Mean & SMD

Table 5 presents the average human and machine mean scores and SMDs for each scoring method for the test set (20% of the whole data). Human mean scores differed slightly for each scoring method because each method filters out a different number of untranslatable responses. While all four scoring methods exhibited average absolute SMDs below 0.15 across the items, MMT_ANNs and UL demonstrated the smallest average SMDs of 0. This result suggests that, on average, nearly identical mean scores were obtained between human and MMT_ANN scores, and human and UL scores.

Table 5

Average mean scores and SMDs for automated scoring methods

	Google_ANNs			ChatGPT_ANNs			MMT_ANNs			Unsupervised Learning		
	H	M	SMD	H	M	SMD	H	M	SMD	H	M	SMD
Item 1	0.63	0.59	−0.07	0.62	0.66	0.08	0.63	0.64	0.02	0.63	0.62	−0.03
Item 2	0.58	0.57	−0.04	0.59	0.57	−0.02	0.58	0.57	−0.02	0.59	0.60	0.03
Item 3	0.52	0.50	−0.04	0.52	0.54	0.05	0.52	0.52	−0.01	0.52	0.51	−0.01
Average	0.58	0.55	−0.05	0.58	0.59	0.04	0.58	0.58	0.00	0.58	0.58	0.00

Notes. H = Human scoring; M = Automated scoring; SMD = Standardized mean score difference

4.2.3 Agreement Statistics: EA & Kappa

Table 6 displays the average EA and κ statistics for each scoring method for the test set. All four scoring methods exhibited high average EA and κ values across the items, exceeding 90% and 0.75, respectively. MMT_ANNs demonstrated the highest average EA (93.85%) and κ (0.87) values.

Table 6

Average EA and Cohen's kappa for automated scoring methods

	Google_ANNs		ChatGPT_ANNs		MMT_ANNs		Unsupervised Learning	
	EA	κ	EA	κ	EA	κ	EA	κ
Item 1	91.52%	0.81	89.45%	0.77	92.49%	0.83	91.54%	0.81
Item 2	93.69%	0.86	91.34%	0.81	94.19%	0.87	89.64%	0.77
Item 3	93.65%	0.85	91.61%	0.80	94.88%	0.87	91.95%	0.81
Average	92.95%	0.84	90.80%	0.79	93.85%	0.86	91.04%	0.80

Note. EA = Exact agreement

After comparing the multiple scoring methods, we determined that MMT_ANNs consistently demonstrated the best performance. This method yielded the lowest average amount of untranslated responses, the lowest SMD values, and the highest EA and κ values. These values indicate that the MMT_ANNs perform most like human raters for the test set across items. Thus, we proceeded with a comprehensive examination of the MMT_ANNs' results based on the aforementioned evaluation metrics, along with the misalignment analysis between human and automated scoring.

4.3 Comprehensive Examination of MMT_ANNs

4.3.1 MT Performance: Untranslated Responses in MMT

MMT_ANNs effectively translated most non-English responses into English, with less than 1% of responses left untranslated on average across items. MMT_ANNs selected the more suitable MT tool, Google Translate or ChatGPT, depending on the individual language groups and items. This strategic MT selection enhanced translation quality and yielded more accurate BOW feature representation. Table 7 presents the number and percentage of non-English languages that used Google Translate or ChatGPT as the ultimate MT engine. On average, approximately 70% of languages were assigned Google Translate as the final MT, whereas the remaining 30% were assigned ChatGPT, often including LRLs such as Armenian, Basque, Galician, or Macedonian.

Table 7

Number and percentage of final MT engines

	Google Translate		ChatGPT		Total	
	n	%	*n*	%	*n*	%
Item 1	27	65.85%	14	34.15%	41	100%
Item 2	30	73.17%	11	26.83%	41	100%
Item 3	21	72.41%	8	27.59%	29	100%
Average	26	70.48%	11	29.52%	37	100%

Furthermore, fine-tuning translations at the individual response level improved MT performance by reducing the percentage of untranslated responses. This adjustment was particularly beneficial for translating misspelled responses and LRL responses. For example, for Item 1, human raters scored the responses in Table 8—Chinese, Turkish, Azerbaijani, and Armenian—as correct. The first two responses included spelling errors (according to the native speakers at the TIMSS & PIRLS International Study Center), and the last two were in LRL. Although Google Translate was initially selected as the MT option for the first three languages (Chinese, Turkish, and Armenian), MMT_ANNs transitioned from Google to ChatGPT translation for these responses due to inappropriate translations where Google printed transliteration or original language responses instead of English. Also, despite ChatGPT being the primary MT tool for Azerbaijani, a few responses, like the last one in Table 8, utilized Google Translate when ChatGPT flagged it as "uncertain." The adjustment percentage for individual responses was notable for LRLs: 15% (Item 1) and 9.34% (Item 3) of ChatGPT-translated responses in Azerbaijani were substituted with Google translations. Similarly, for Albanian and Kazakh, 8.00% (Item 1) and 6.68% (Item 1) of Google-translated responses were replaced with ChatGPT-translated responses.

Table 8

Translation adjustment for misspelled or LRL responses

Language	Original Response	Google Translate	ChatGPT	MMT_ANNs
Chinese	地新影力	Dixin	Gravity	Gravity

		influence		
Turkish	REÇEKİMİ KUVETİ	Cuvet review	force of gravity	force of gravity
Azerbaijani	agirliq quvvesi	agirliq quvvesi	gravitational force	gravitational force
Armenian	ԾԱՆՐՈՒԹՅԱՆ ՈՒԺ ԿԱՄ ՏԻԵԶԵՐԱԿԱՆ ՁԳՈՂՈՒԹՅԱՆ ՈՒԺ	Gravity or cosmic gravity	Uncertain	Gravity or cosmic gravity

4.3.2 Distributional Differences: Mean & SMD

Figures 1-3 illustrate the relationship between human and machine mean scores for all countries. Dots below the red diagonal indicate that human mean scores are higher than machine mean scores, while dots above the diagonal indicate the opposite. Across the three items, most dots are clustered around the diagonal, indicating that human and machine mean scores were generally very similar. High Pearson correlations were observed across all three items (r_{item1} = 0.82; r_{item2} = 0.98; r_{item3} = 0.98). However, deviations were noted in several instances, such as C17, C27, and C33 for Item 1, C58 in Item 2, and C14 and C23 for Item 3. These deviations were also reflected in the SMDs. Of the items assessed, absolute SMDs exceeded 0.15 in eight countries (15.38%) for Item 1, two countries (3.85%) for Item 2, and four countries (10%) for Item 3. These deviations are further described in section 4.4 Misalignment Analysis.

Figure. 1.

Distribution of human-machine mean scores for Item 1

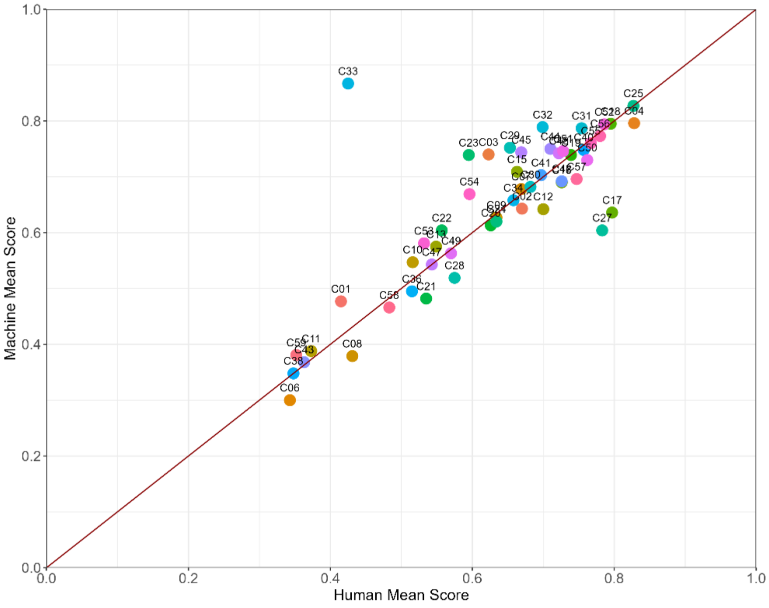


Figure. 2.

Distribution of human-machine mean scores for Item 2

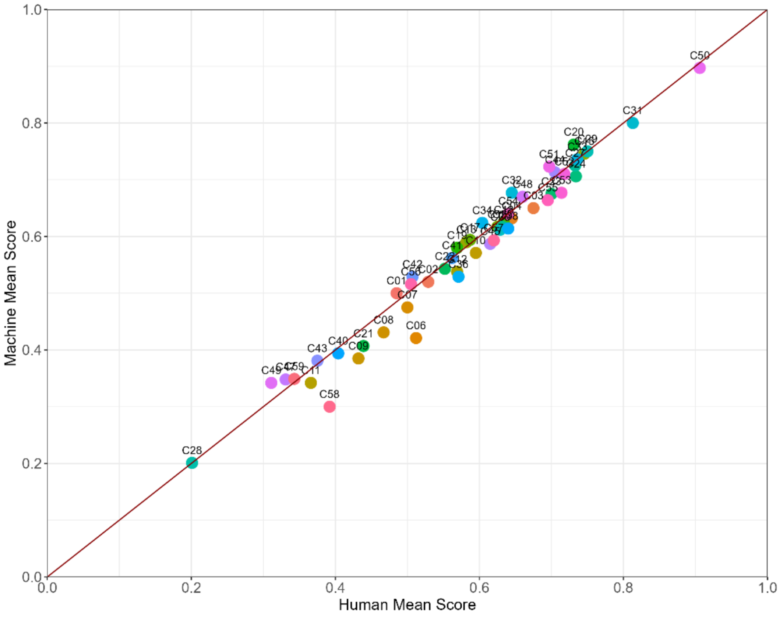
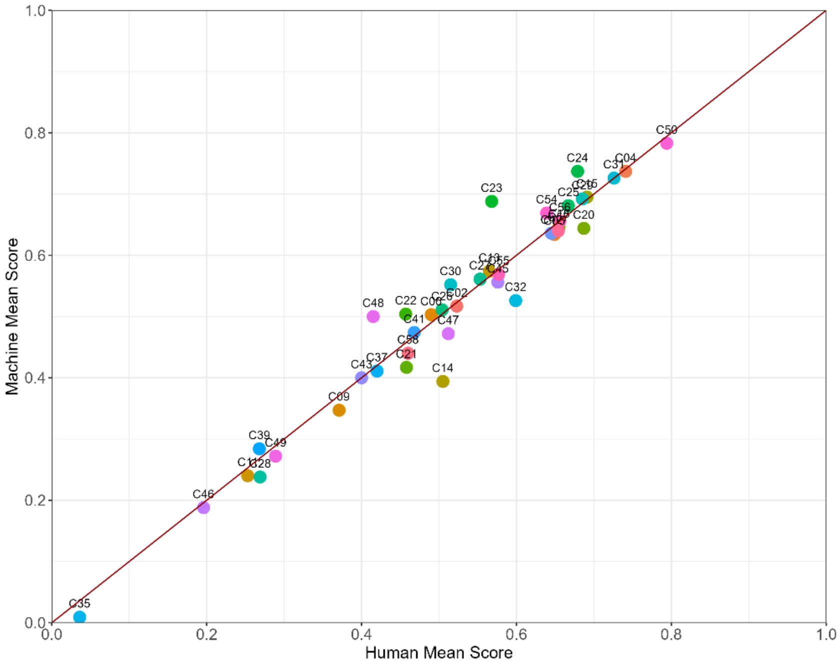


Figure. 3.

Distribution of human-machine mean scores for Item 3



4.3.3 Agreement Statistics: EA & Kappa

Table 9 shows the country-by-item EA and κ values. While the majority of values fell within the preferred range, some instances of moderate EAs were noted in C23 for Item 1 (80.18%) and C03 for Item 2 (79.17%). C33 for Item 1 had a notably low EA of 54.17% and a poor κ value of 0.17. C35 had a κ value of almost zero for Item 3 (κ = −0.01) with an EA of 95.58%, indicating that the high agreement between the human rater and machine classifications was by chance. The poor agreements are further described in the following section (4.4 Misalignment Analysis).

Table 9

Item-by-Country EA & Cohen's kappa from MMT_ANNs

	EA			κ		
Country	Item 1	Item 2	Item 3	Item 1	Item 2	Item 3
C01	84.62%	94.12%		0.69	0.88	
C02	95.55%	93.91%	94.65%	0.90	0.88	0.89
C03	80.52%	79.17%		0.56	0.54	
C04	96.82%	94.19%	95.47%	0.90	0.87	0.88
C05			95.42%			0.90
C06	92.86%	81.71%	93.10%	0.84	0.64	0.86
C07	94.21%	96.72%		0.87	0.93	
C08	93.10%	95.62%		0.86	0.91	
C09	97.89%	92.57%	95.29%	0.95	0.85	0.90
C10	96.88%	96.43%		0.94	0.93	
C11	94.90%	91.46%	97.39%	0.89	0.81	0.93
C12	91.51%	93.77%		0.81	0.87	

C13	95.58%	93.44%	99.07%	0.91	0.87	0.98
C14			81.65%			0.63
C15	95.43%	96.83%	96.82%	0.89	0.92	0.93
C16	94.69%	97.50%		0.87	0.95	
C17	82.52%	94.20%		0.58	0.88	
C18	98.21%	95.54%	97.35%	0.95	0.91	0.94
C19	96.44%	92.03%		0.91	0.84	
C20	94.84%	98.12%	95.71%	0.89	0.95	0.90
C21	94.74%	91.87%	95.83%	0.89	0.83	0.92
C22	93.40%	94.02%	88.98%	0.86	0.88	0.78
C23	80.18%	96.67%	86.40%	0.56	0.92	0.71
C24	98.59%	95.80%	92.95%	0.97	0.90	0.83
C25	97.74%	96.99%	98.58%	0.92	0.94	0.97
C26			96.40%			0.93
C27	82.08%	98.33%	99.19%	0.59	0.96	0.98
C28	89.44%	95.10%	90.77%	0.79	0.85	0.76

C29	90.10%	96.88%	95.21%	0.77	0.92	0.89
C30	95.45%	89.89%	89.09%	0.90	0.79	0.78
C31	96.75%	96.00%	98.86%	0.91	0.87	0.97
C32	87.80%	94.35%	91.45%	0.68	0.87	0.83
C33	54.17%	97.52%		0.17	0.94	
C34	92.11%	92.62%		0.82	0.84	
C35			95.58%			−0.01
C36	95.96%	92.44%		0.92	0.85	
C37			99.11%			0.98
C38	95.51%	97.37%		0.90	0.94	
C39			96.80%			0.92
C40	93.20%	98.99%		0.82	0.98	
C41	97.93%	97.92%	97.08%	0.95	0.96	0.94
C42	96.58%	93.43%	97.27%	0.92	0.87	0.94
C43	93.53%	94.10%	97.04%	0.86	0.87	0.94
C44	95.97%	94.57%		0.90	0.87	

C45	92.48%	92.31%	97.92%	0.82	0.84	0.96
C46			93.81%			0.80
C47	93.14%	95.58%	94.31%	0.86	0.90	0.89
C48	97.94%	94.85%	91.51%	0.95	0.89	0.83
C49	94.04%	93.69%	98.25%	0.88	0.85	0.96
C50	94.71%	97.44%	98.94%	0.86	0.87	0.97
C51	91.23%	95.80%		0.77	0.90	
C52	97.71%	96.95%		0.93	0.93	
C53	87.10%	96.99%		0.74	0.93	
C54	91.18%	95.68%	86.47%	0.81	0.91	0.70
C55	97.67%	96.09%	99.23%	0.94	0.91	0.98
C56	98.00%	91.21%	98.19%	0.94	0.82	0.96
C57	93.68%	92.25%	96.05%	0.84	0.84	0.91
C58	93.22%	83.33%	92.00%	0.86	0.63	0.84
C59	89.52%	93.60%		0.77	0.86	
Average	92.49%	94.19%	94.88%	0.83	0.87	0.87

4.4 Misalignment Analysis

The sources of misalignment between human and automated scoring were varied (see Table 10). We analyzed instances where at least one of the metrics (i.e., mean score, SMD, EA, κ) flagged performance issues in MMT_ANNs compared to human scoring.

First, we noticed that human scoring subjectivity or inconsistency is one crucial cause of the misalignment between human and automated scoring. According to the consultation with a native speaker at the ISC regarding C33 responses to Item 1 (EA = 54.17%, κ = 0.17), we learned that severe scoring contributed to the large discrepancy when comparing human and machine scores. It appears that the human raters scored responses containing outdated terms or indirect explanations of keywords as incorrect. In contrast, MMT_ANNs consistently classified these responses as correct. These responses accounted for 12 (22%) out of 55 misclassified responses in the test set.

Next, we observed human scoring inconsistency even across the same language in different countries. For instance, for Item 1, human scoring displayed inconsistencies in evaluating identical Russian misspelled responses across five Russian-speaking countries: one country scored them as incorrect, while the others considered them as correct. Similarly, within three Chinese-speaking countries, discrepancies arose in scoring the same Chinese response that was a synonym for the keyword to Item 1; two countries marked this variation on the keyword as correct, whereas one country tended to not give the same response credit. This country was C23, which showed moderate EA (EA = 80.18%, κ = 0.56); 18 (83%) out of 22 misclassified responses had been scored differently than in the other two Chinese speaking countries. While the exact reason for the scoring discrepancy is not known, there is the possibility that the response was included in the translation of the scoring guide for the other two countries but not C23, or potentially that the phrase was not familiar to the human raters in C23.

Translation issues from MT were another source of misalignment. The causes of mistranslation were mainly related to (1) a variety of misspellings, (2) LRL responses, and (3) polysemy. First, many misspelled and non-existing words were an important cause of translation issues. Misspelled responses are often phonetically similar to correctly spelled words but are lexically different. As a result, Google Translate and ChatGPT usually provided incorrect translations for heavily misspelled responses. For example, for Item 1, 25 (96%) out of 26 misclassified responses in C17 contained significant misspellings. Our fuzzy matching approach did not address these misspelled responses because they deviated even more than two edit distances from the responses considered correct. ChatGPT occasionally generated hallucinated outputs when these responses were untranslatable. Hallucinations are realistic outputs generated by a machine, such as ChatGPT, that do not correspond to real-world input (Alkaissi & McFarlane, 2023). ChatGPT tended to produce hallucinations for significantly misspelled or unrecognizable LRL responses, which caused the "translated" responses to appear as correct answers regardless of the original response. For instance, for Item 1, 12 (80%) out of 15 misclassified responses in C03 and all 10 misclassified responses in C45 resulted from hallucinations, where the discrepancy in classifications came from human raters scoring the responses as incorrect but the ANNs giving them credit. Yet, given that Google Translate also struggled to translate these hallucinated cases, it is presumed that the translation difficulty might arise from misspellings or the challenge of translating LRL responses. Also, Google Translate sometimes provided another possible but context-irrelevant translation for keywords in cases where the original language had multiple possible meanings (i.e., polysemy).

Table 10

Sources of misalignment between human and automated scoring

Source	Item	Country	Language	Note
Human Scoring Issue	1	C23		A human rater scored certain responses differently from those in other same-language countries.
		C33	LRL	A human rater scored older terms or indirect explanations of keywords as incorrect. A human rater scored certain responses differently from those in other same-language countries.
	3	C23		A human rater scored certain responses differently from those in other same-language countries.
		C48	LRL	A human rater scored lexically similar but misspelled responses as incorrect.
Automated Scoring Issue	1	C03	LRL	ChatGPT provided hallucinated outputs for incorrect responses.
		C17		Both Google and ChatGPT struggled to translate certain heavily misspelled responses.
		C27		Google provided an alternative possible meaning for certain responses.

		C29		ChatGPT provided hallucinated translations for incorrect responses.
		C45	LRL	ChatGPT provided hallucinated translations for incorrect responses.
	2	C06	LRL	ChatGPT struggled to translate a certain word (keyword) into English.
		C58	LRL	Google struggled to translate a certain word (keyword) into English.
	3	C14	LRL	ChatGPT struggled to translate a certain word (keyword) into English.
Automated & Human Scoring Issue	1	C32	LRL	A human rater scored lexically similar but misspelled keywords as incorrect.
	2	C03	LRL	Human raters scored some off-topic responses as correct and human-human agreement was notably low (EA = 76.40%, κ = 0.39). Both Google and ChatGPT struggled to translate some of the responses.
Other	3	C35	LRL	An extremely skewed distribution (low performance) was observed in the very small test set (n=88).

Note. LRL = Low-resource language

5. Discussion

5.1 The Potential of Automated Scoring in ILSAs

The results suggest that automated scoring performs similarly to human scoring and demonstrates the potential of applying automated scoring of short CR items in ILSAs using MT. Specifically, MMT_ANNs consistently demonstrated highly accurate scoring across various items and countries, surpassing other automated scoring methods. The impressive performance of MMT_ANNs can be attributed to their adaptability in selecting more suitable translations at both the individual language and response levels. Our MMT method extended previous studies who found that combining multiple MT techniques could improve translation quality (Banik et al., 2019; Costa-Jussa & Fonollosa, 2015). The need to integrate various MTs arises from the limitation of a single MT to achieve the desired level of accuracy (Kahlon & Singh, 2023). In line with prior research, our MMT approach generally produced accurate translations for most languages used in the TIMSS 2023 dataset and wasparticularly useful for addressing translation challenges in LRL responses. Moreover, the MMT method was accurate enough to generate a BOW feature matrix that could be used to score responses across 42 different languages and 52 countries.

Automated scoring can handle the two primary challenges associated with human scoring: (1) subjectivity and inconsistency and (2) high costs. First, automated scoring enhances scoring consistency by applying the same scoring criteria to all responses. We observed inconsistencies in human scoring within and between countries, particularly for borderline responses that fall between "definitely correct" and "definitely incorrect" (Mitchell et al., 2002). Some borderline responses, such as older terms or heavily misspelled answers, were scored inaccurately or inconsistently by human raters. Even within countries sharing the same language, identical responses were occasionally assigned different scores. This inconsistency arises because, in practice, it is highly challenging for human raters to reach a unified scoring standard due to individual rater idiosyncrasies (Attali, 2014; Bejar, 2012). Moreover, TIMSS scoring guide translation is conducted by national representatives on a country-by-country basis, where two countries that share the same language could yield slightly different scoring guides and deliver different training to their human raters. In contrast, automated scoring classifies all responses consistently, including borderline cases, across countries and languages. There is also the potential that because automated scoring applies the same rules to all responses, it could help identify scoring differences across cycles if responses from past cycles were included. Also, automated scoring substantially reduces the workload and costs associated with scoring large volumes of responses in ILSAs. With the availability of low-cost computing resources, automated scoring could be a resource-efficient approach to enhance scoring reliability, reducing the necessity for hiring second human raters.

In summary, automated scoring can be a valuable and affordable additional quality control measure in ILSAs. Previous studies have shown that automated scoring can be used to monitor the quality and consistency of CR scoring (Shaw et al., 2020; Wang & von Davier, 2014; Williamson, 2012). Alongside traditional psychometric statistics such as item difficulty, discrimination and differential item functioning, human-machine agreement statistics can be employed to monitor reliable scoring. Large discrepancies between human and automated scoring may offer valuable information for both types of scoring that would not have been identified otherwise. For instance, C33 had unusually low performance for Item 1 compared to other countries, where the percent of students answering the item correctly was much lower than the international average. The current inter-rater reliability evidence shows no apparent scoring issues (human-human agreement = 96.50%). However, this study found that C33 had unusually low human-machine agreement, prompting further investigation. This review, which may not have been conducted otherwise, revealed scoring that was too strict. Therefore, the automated scoring validation not only pointed to a possible reason for the item's unusual performance in C33, but it also reduced the amount of data that would need to be reviewed by an expert (i.e., providing only responses where the human rater and machine classifications did not match).

Discrepancies in human-machine classifications could stem from human scoring issues due to rater effects or from translation issues. Further investigation of these discrepancies would be beneficial for either case. For human scoring issues, we might review the scoring guide to check if scoring for certain responses may be unclear. If problems with the individual countries' scoring is suspected , country representatives could be asked to look at the data and reach out to their scoring team with questions. Regarding translation issues, examining the probable causes of mis- or untranslated responses is crucial to refine the translation process in automated scoring for ILSAs. After fully implementing automated scoring for quality control, we can proceed to prepare for operational use by revisiting the methodologies.

Despite all the advantages, it is crucial to acknowledge that the performance of automated scoring relies on the quality of MT. Even with MMT, translations may still be improper or untranslatable depending on the extent of lexical variations (e.g., spelling errors) or the characteristics of the language. While human raters scored various misspelled responses as correct based on their interpretation of the intended words, automated scoring could classify them as incorrect due to inaccurate translations resulting from spelling errors. MT appears to have more difficulty translating heavily misspelled responses, especially for LRLs. Such inaccurate translations can introduce bias in favor or against certain groups or languages.

However, with significant advances in NLP and LLM, we anticipate improvements in spelling correction and translation accuracy within the next year or so. The current race of big-tech companies and the grassroots activities around releasing new LLMs, on the one hand, and finetuning and adapting for specific tasks, on the other hand, will likely yield more capable models suitable for MT to and from virtually all languages represented online. It is expected that potential improvements will help minimize bias in automated scoring. Furthermore, since tolerance for misspelled responses may vary among human raters, there is also potential for bias in human scoring due to misspellings that are treated differently from scorer to scorer. Therefore, automated scoring should focus on consistent and accurate assessment within and across countries rather than simply emulating human scoring.

5.2 Limitations

This study had several notable limitations. One limitation is that MMT_ANNs were based on the BOW, which ignores word order and context. The BOW approach limited which items could be selected for automated scoring: items with extended responses and those where the same words could be used in both correct and incorrect responses would not have been scored accurately. Another limitation is that the performance of ANNs relies heavily on the quality of the training set. If the training set contains improper human scoring or translation issues, it becomes challenging to build an accurate model (Chollet, 2021). Ideally, the training set should be cleaned before modeling. One option could be to use UL to classify responses in the training set and only use responses where both the human rater and UL scores match since they are more likely to be scored properly.

5.3 Future Research

Future researchers could explore alternative NLP techniques and sampling methodology. First, sentence embedding could be investigated as an alternative to the BOW approach. Unlike the BOW, sentence embedding incorporates context and retains the overall meaning of responses, which could improve ANN accuracy by retaining more of students' written text. This approach could also expand the pool of items that can be validated to include CR items that elicit longer responses or responses with multiple synonyms. Another potential area of exploration could be data augmentation and varying the sample assignment distribution. In our study, only 20% of responses were used in the test set. However, augmenting existing student responses or using LLMs to provide additional student responses for training could increase the number of responses in the test set and, thus, the number of responses that could be double-scored operationally in ILSAs.

Future research should also consider expanding the automated scoring methods by further investigating UL and/or incorporating multiple automated scoring engines into the validation process. First, our results suggested that while MMT_ANNs outperformed UL, UL still holds promise as a fast way to generate classifications without the need for training on human rater scores. UL demonstrated good performance across countries and languages and warrants further investigation. The approach could be particularly valuable in cases where potential scoring or translation issues exist within the training set. Consideration should also be given to expanding the validation method past a second set of classifications, instead generating three or more sets of classifications using different scoring engines. Recently, Verga et al. (2024) explored evaluating automated scoring by comparing scores generated by a panel of small LLMs. They found that multiple sets of classifications by smaller models could be used in place of a single set of classifications from a large model (such as GPT-4) because it reduces bias in evaluation (e.g., from prompt wording, formatting, hallucinations, etc.). There is the potential multiple scoring engine classifications could help narrow down which item-country groups warranted further investigation, where only those with large disparities in agreement across most engines would be reviewed by experts.

6. Conclusion

This study provides evidences for the possible use of automated scoring as a quality control measure in ILSAs. Automated scoring could enhance scoring consistency, accuracy, and cost-effectiveness while mitigating the challenges associated with human scoring. Leveraging state-of-the-art MT contributed to building a unified scoring model that can reliably score responses across languages and countries. Because ILSA results inform educational policy changes for participating countries, it is vital that the measure be as accurate as possible. Ultimately, improving CR scoring in ILSAs through AI-based validation will lead to more precise changes in country-level education systems, and thus better student learning in the future.

References

Alkaissi, H., & McFarlane, S. I. (2023). Artificial hallucinations in ChatGPT: implications in scientific writing. *Cureus*, *15*(2). https://doi.org/10.7759/cureus.35179

Araújo, M., Pereira, A., & Benevenuto, F. (2020). A comparative study of machine translation for multilingual sentence-level sentiment analysis. *Information Sciences*, *512*, 1078-1102. https://doi.org/10.1016/j.ins.2019.10.031

Arun, K., & Srinagesh, A. (2020). Multilingual Twitter sentiment analysis using machine learning. *International Journal of Electrical and Computer Engineering*, *10*(6), 5992-6000. http://doi.org/10.11591/ijece.v10i6.pp5992-6000

Attali, Y. (2014). A ranking method for evaluating constructed responses. *Educational and Psychological Measurement*, *74*(5), 795-808. https://doi.org/10.1177/0013164414527450

Azaria, A., Azoulay, R., & Reches, S. (2024). ChatGPT is a remarkable tool—for experts. *Data Intelligence*, *6*(1), 240-296. https://doi.org/10.1162/dint_a_00235

Banik, D., Ekbal, A., Bhattacharyya, P., & Bhattacharyya, S. (2019). Assembling translations from multi-engine machine translation outputs. *Applied Soft Computing*, *78*, 230-239. https://doi.org/10.1016/j.asoc.2019.02.031

Bejar, I. I. (2012). Rater cognition: implications for validity. *Educational Measurement: Issues and Practice*, *31*(3), 2-9. https://doi.org/10.1111/j.1745-3992.2012.00238.x

Bennett, R. E. (1991). On the meanings of constructed response. *ETS Research Report Series*, *1991*(2), i-46. https://doi.org/10.1002/j.2333-8504.1991.tb01429.x

Braun, H. I., Bennett, R. E., Frye, D., & Soloway, E. (1990). Scoring constructed responses using expert systems. *Journal of Educational Measurement*, *27*(2), 93-108. https://doi.org/10.1111/j.1745-3984.1990.tb00736.x

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

Cayrol, M., Farreny, H., & Prade, H. (1982). Fuzzy pattern matching. *Kybernetes*, *11*(2), 103-116. https://doi.org/10.1108/eb005612

Choi, H., Cho, K., & Bengio, Y. (2018). Fine-grained attention mechanism for neural machine translation. *Neurocomputing*, *284*, 171-176. https://doi.org/10.1016/j.neucom.2018.01.007

Chollet, F. (2021). *Deep learning with Python*. Simon and Schuster.

Cieri, C., Maxwell, M., Strassel, S., & Tracey, J. (2016, May). Selection criteria for low resource language programs. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (pp. 4543-4549). https://aclanthology.org/L16-1720.pdf

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

Costa-Jussa, M. R., & Fonollosa, J. A. (2015). Latest trends in hybrid machine translation and its applications. *Computer Speech & Language*, *32*(1), 3-10. https://doi.org/10.1016/j.csl.2014.11.001

Dubois, D., Prade, H., & Testemale, C. (1993). Weighted fuzzy pattern matching. In *Readings in Fuzzy Sets for Intelligent Systems* (pp. 676-685). Morgan Kaufmann. https://doi.org/10.1016/B978-1-4832-1450-4.50073-0

El Naqa, I., & Murphy, M. J. (2015). *What is machine learning?* (pp. 3-11). Springer International Publishing. https://link.springer.com/chapter/10.1007/978-3-319-18305-3_1

Federer, M. R., Nehm, R. H., Opfer, J. E., & Pearl, D. (2015). Using a constructed-response instrument to explore the effects of item position and item features on the assessment of students' written scientific explanations. *Research in Science Education*, *45*, 527-553. https://doi.org/10.1007/s11165-014-9435-9

Gilson, A., Safranek, C. W., Huang, T., Socrates, V., Chi, L., Taylor, R. A., & Chartash, D. (2023). How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. *JMIR Medical Education*, *9*(1), e45312. http://doi.org/10.2196/45312

Horbach, A., Pehlke, J., Laarmann-Quante, R., & Ding, Y. (2023). Crosslingual content scoring in five languages using machine-translation and multilingual transformer models. *International Journal of Artificial Intelligence in Education*, 1-27. https://doi.org/10.1007/s40593-023-00370-1

Jiao, W., Wang, W., Huang, J. T., Wang, X., & Tu, Z. (2023). *Is ChatGPT a good translator? A preliminary study.* arXiv preprint. https://doi.org/10.48550/arXiv.2301.08745

Johnson, M. S., & McCaffrey, D. F. (2023). Evaluating fairness of automated scoring in educational measurement. In Yaneva, V., & von Davier, M. (Eds.) *Advancing natural language processing in educational assessment* (pp. 142-163). Routledge.

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

Jung, J. Y., Tyack, L., & von Davier, M. (2024). Combining machine translation and automated scoring in international large-scale assessments. *Large-scale Assessments in Education*, *12*, 10. https://doi.org/10.1186/s40536-024-00199-7

Kahlon, N. K., & Singh, W. (2023). Machine translation from text to sign language: a systematic review. *Universal Access in the Information Society*, *22*(1), https://doi.org/10.1007/s10209-021-00823-1

Kim, Y. (2014). *Convolutional neural networks for sentence classification*. arXiv preprint. 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

LaVoie, N., Parker, J., Legree, P. J., Ardison, S., & Kilcullen, R. N. (2020). Using latent semantic analysis to score short answer constructed responses: automated scoring of the consequences test. *Educational and Psychological Measurement*, *80*(2), 399-414. https://doi.org/10.1177/0013164419860575

Levenshtein, V. I. (1966, February). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, *10*, 707-710. https://nymity.ch/sybilhunting/pdf/Levenshtein1966a.pdf

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

Liu, O. L., Lee, H. S., & Linn, M. C. (2011). Measuring knowledge integration: validation of four‐year assessments. *Journal of Research in Science Teaching*, *48*(9), 1079-1107. https://doi.org/10.1002/tea.20441

Liu, O. L., Rios, J. A., Heilman, M., Gerard, L., & Linn, M. C. (2016). Validation of automated scoring of science assessments. *Journal of Research in Science Teaching*, *53*(2), 215-233. https://doi.org/10.1002/tea.21299

Livingston, S. A. (2009). *Constructed-response test questions: Why we use them; How we score them*. (Research Report No. RDC-11). Retrieved from https://files.eric.ed.gov/fulltext/ED507802.pdf

Lottridge, S., Wood, S., & Shaw, D. (2018). The effectiveness of machine score-ability ratings in predicting automated scoring performance. *Applied Measurement in Education*, *31*(3), 215-232. https://doi.org/10.1080/08957347.2018.1464452

Luong, M. T., Pham, H., & Manning, C. D. (2015). *Effective approaches to attention-based neural machine translation*. arXiv preprint. https://doi.org/10.48550/arXiv.1508.04025

Mao, L., Liu, O. L., Roohr, K., Belur, V., Mulholland, M., Lee, H. S., & Pallant, A. (2018). Validation of automated scoring for a formative assessment that employs scientific argumentation. *Educational Assessment*, *23*(2), 121-138. https://doi.org/10.1080/10627197.2018.1427570

Martin, M. O., von Davier, M., & Mullis, I. V. S. (Eds.). (2020). *Methods and Procedures: TIMSS 2019 Technical Report*. Retrieved from Boston College, TIMSS & PIRLS International Study Center website: https://timssandpirls.bc.edu/timss2019/methods

McCarthy, E. D. (2005). *Knowledge as culture: the new sociology of knowledge*. Routledge.

McClellan, C. A. (2010). *Constructed-response scoring—Doing it right.* (Research Report No. RDC-13). Retrieved from https://www.ets.org/Media/Research/pdf/RD_Connections13.pdf

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/.

Mielke, S. J., Cotterell, R., Gorman, K., Roark, B., & Eisner, J. (2019). *What kind of language is hard to language-model?.* arXiv preprint. https://doi.org/10.48550/arXiv.1906.04726

Mitchell, T., Russell, T., Broomhead, P., & Aldridge, N. (2002). Towards robust computerised marking of free-text responses. https://hdl.handle.net/2134/1884

Mullis, I.V.S, Martin, M.O., & von Davier, M. (Eds.). (2021). *TIMSS 2023 Assessment Frameworks*. Retrieved from Boston College, TIMSS & PIRLS International Study Center website: https://timssandpirls.bc.edu/timss2023

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. https://jimelwood.net/students/grips/tables_figures/myford_wolfe_2004.pdf

OECD (2024). PISA 2025 Science Framework. https://pisa-framework.oecd.org/science-2025/assets/docs/PISA_2025_Science_Framework.pdf

Okubo, T., et al. (2023), AI scoring for international large-scale assessments using a deep learning model and multilingual data, *OECD Education Working Papers*, No. 287, OECD Publishing, Paris, https://doi.org/10.1787/9918e1fb-en.

Prabhakar, D. K., & Pal, S. (2018). Machine transliteration and transliterated text retrieval: a survey. *Sādhanā*, 43(6), 93. https://www.ias.ac.in/public/Volumes/sadh/043/06/0093.pdf

Salminen, J., Kamel, A. M., Jung, S. G., & Jansen, B. (2021). *The problem of majority voting in crowdsourcing with binary classes*. The 19th European Conference on Computer-Supported Cooperative Work, Remote via Internet & Zurich, Switzerland. https://dl.eusset.eu/bitstream/20.500.12015/4163/1/ecscw2021-n12.pdf

Sazli, M. H. (2006). A brief review of feed-forward neural networks. *Communications Faculty of Sciences University of Ankara Series A2-A3 Physical Sciences and Engineering*, *50*(01). https://doi.org/10.1501/commua1-2_0000000026

Shaw, D., Bolender, B., & Meisner, R. (2020). Quality control for automated scoring in large-scale assessment. In Yan, D., Rupp, A. A., & Foltz, P. W. (Eds.). *Handbook of automated scoring: theory into practice* (pp. 241-262). Chapman and Hall/CRC.

Shermis, M. D., & Wilson, J. (Eds.). (2024). *The Routledge international handbook of automated essay evaluation*. Taylor & Francis.

Shermis, M., & Lottridge, S. (2019, April). *Communicating to the public about machine scoring: What works, what doesn't*. National Conference on Measurement in Education (NCME), Toronto, CA. https://www.cambiumassessment.com/-/media/project/cambium/corporate/pdfs/cai-cambium-communicatingpublicmachinescoring-whitepaper.pdf

Shin, H. J., Andersen, N., Horbach, A., Kim, E., Baik, J., & Zehner, F. (2024). Operational automatic scoring of text responses in 2016 ePIRLS: performance and linguistic variance. https://www.iea.nl/sites/default/files/2024-04/Operational-Automatic-Scoring-of-Text-Responses-ePIRLS.pdf

Sukkarieh, J. Z., & Blackmore, J. (2009, March). *C-rater: Automatic content scoring for short constructed responses*. The Florida Artificial Intelligence Research Society (FLAIRS), Sarasota, FL. https://cdn.aaai.org/ocs/122/122-2394-1-PB.pdf

Surden, H. (2021). Machine learning and law: An overview. In Vogl, D. (Ed.) *Research Handbook on Big Data Law* (pp. 171-184), Edward Elgar Publishing. https://doi.org/10.4337/9781788972826

Tu, Z., Lu, Z., Liu, Y., Liu, X., & Li, H. (2016). *Modeling coverage for neural machine translation*. arXiv preprint. https://doi.org/10.48550/arXiv.1601.04811

Verga, P., Hofstatter, S., Althammer, S., Su, Y., Piktus, A., Arkhangorodsky, A., Xu, M., White, N. and Lewis, P. (2024). Replacing Judges with Juries: Evaluating LLM Generations with a Panel of Diverse Models. *arXiv preprint arXiv:2404.18796*. https://arxiv.org/abs/2404.18796

Wang, Z., & von Davier, A. A. (2014). Monitoring of scoring using the e‐rater® automated scoring system and human raters on a writing test. *ETS Research Report Series*, *2014*(1), 1-21.

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

Zhang, M. (2013). *Contrasting automated and human scoring of essays.* (Research Report No. RDC-21). Retrieved from https://www.ets.org/Media/Research/pdf/RD_Connections_21.pdf

CHAPTER 4. CONCLUSION

The three dissertation papers on automated scoring present a series of increasingly more adaptable solutions to address the challenge of handling the multitude of multilingual responses. We aim to incorporate AI innovations such as machine translation and large language models to build an automated scoring model for multilingual responses. In contrast to automated scoring solely in English or high-resource languages, the scoring method outlined in this dissertation offers feasible solutions even for low-resource languages where advanced NLP techniques may not be accessible. These proposed solutions establish a viable workflow, including assessments of machine translation quality and scoring validity through various metrics like human-machine score agreements, standard mean score differences, and kappa statistics. Investigating comparability with human scoring across different languages and regions ensures that the proposed methods produce scores that are comparable to or even more accurate than human scores. This potentially enhances the psychometric properties for subsequent analyses. Furthermore, the implementation of automated scoring can be a critical quality control measure for ILSAs, while significantly reducing reliance on secondary human raters.