Psychological Test and Assessment Modeling, Volume 64, 2022 (4), 471-494

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

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

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.

Keywords: International large-scale assessment, eTIMSS, constructed-response items, automated scoring, artificial neural networks, natural language processing

[1] TIMSS & PIRLS International Study Center at Boston College *Correspondence concerning this article should be addressed to:* Ji Yoon Jung, TIMSS & PIRLS International Study Center at Boston College, 140 Commonwealth Ave. Chestnut Hill, MA 02467, USA. jiyoon.jung@bc.edu

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

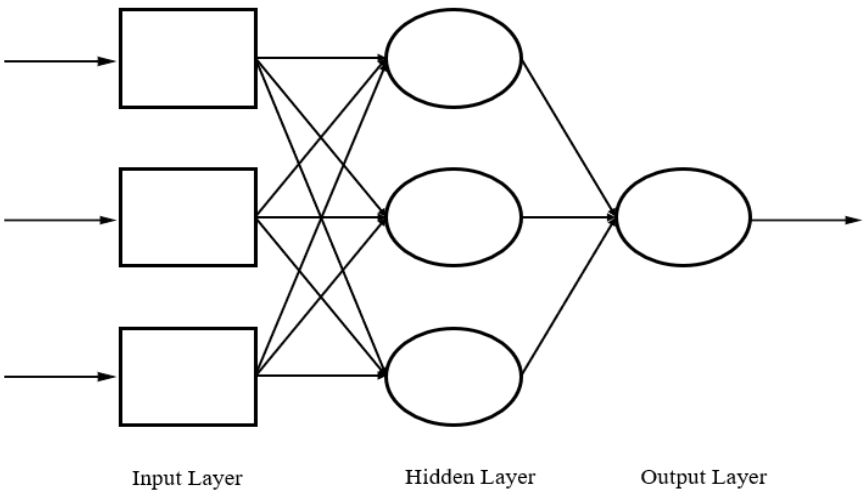
responses (N = 1,863) written by Reserve Officers' Training Corps cadets from the Consequences Test, a measure of creativity and divergent thinking. Automated scores demonstrated very high convergence with human raters (r = 0.94) and provided similar patterns of predictive and concurrent validity as 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.

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

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).

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

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.

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

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://timss-sandpirls.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