Automated Scoring of Open-Ended Situational Judgment Tests

Authors: Xu Jing, Luo Fang, Ma Yanzhen, Hu Luming, Tian Xuetao, Tian Xuetao

Date: 2023-12-21T00:00:00+00:00

Abstract

Due to scoring cost limitations, open-ended situational judgment tests are difficult to widely use. This study explores the application of automated scoring using teacher competency assessment as an example. An open-ended situational judgment test was developed for typical problem scenarios in teaching, and response texts from primary and secondary school teachers were collected. A supervised learning strategy was employed to apply deep neural networks at both the document level and sentence level to identify response categories. Convolutional Neural Network (CNN) achieved ideal results, with scoring accuracy ranging from 70% to 88% for each item, showing high consistency with human scoring. The correlation coefficient between human and machine scoring was r = 0.95, and the Quadratic Weighted Kappa (QWK) coefficient was 0.82. The results indicate that machine scoring can achieve stable performance, and automated scoring research can facilitate the widespread application of open-ended situational judgment tests.

Full Text

Preamble

Automated Scoring of Open-ended Situational Judgment Tests

XU Jing[1], LUO Fang[1], MA Yanzhen[2], HU Luming[3], TIAN Xuetao[1]
([1] School of Psychology, Beijing Normal University, Beijing 100875, China)
([2] Collaborative Innovation Center of Assessment toward Basic Education Quality, Beijing Normal University, Beijing 100875, China)
([3] Department of Psychology, School of Arts and Sciences, Beijing Normal University at Zhuhai, Zhuhai 519085, China)

Abstract

Situational Judgment Tests (SJTs) have gained popularity for their unique testing content and high face validity. However, traditional SJT formats, particularly those employing multiple-choice (MC) options, have encountered scrutiny due to their susceptibility to test-taking strategies. In contrast, open-ended and constructed response (CR) formats present a propitious means to address this issue. Nevertheless, their extensive adoption encounters hurdles primarily stemming from the financial implications associated with manual scoring. In response to this challenge, we propose an open-ended SJT employing a written-constructed response format for the assessment of teacher competency. This study established a scoring framework leveraging natural language processing (NLP) technology to automate the assessment of response texts, subsequently subjecting the system's validity to rigorous evaluation. The study constructed a comprehensive teacher competency model encompassing four distinct dimensions: student-oriented, problem-solving, emotional intelligence, and achievement motivation. Additionally, an open-ended situational judgment test was developed to gauge teachers' aptitude in addressing typical teaching dilemmas.

A dataset comprising responses from 627 primary and secondary school teachers was collected, with manual scoring based on predefined criteria applied to 6,000 response texts from 300 participants. To expedite the scoring process, supervised learning strategies were employed, facilitating the categorization of responses at both the document and sentence levels. Various deep learning models, including the convolutional neural network (CNN), recurrent neural network (RNN), long short-term memory (LSTM), C-LSTM, RNN+attention, and LSTM+attention, were implemented and subsequently compared, thereby assessing the concordance between human and machine scoring. The validity of automatic scoring was also verified.

This study reveals that the open-ended situational judgment test exhibited an impressive Cronbach's alpha coefficient of 0.91 and demonstrated a good fit in the validation factor analysis through the use of Mplus. Criterion-related validity was assessed, revealing significant correlations between test results and various educational facets, including instructional design, classroom evaluation, homework design, job satisfaction, and teaching philosophy. Among the diverse machine scoring models evaluated, CNNs have emerged as the top-performing model, boasting a scoring accuracy ranging from 70% to 88%, coupled with a remarkable degree of consistency with expert scores ($r = 0.95$, QWK = 0.82). The correlation coefficients between human and computer ratings for the four dimensions—student-oriented, problem-solving, emotional intelligence, and achievement motivation—approximated 0.9. Furthermore, the model showcased an elevated level of predictive accuracy when applied to new text datasets, serving as compelling evidence of its robust generalization capabilities.

This study ventured into the realm of automated scoring for open-ended situational judgment tests, employing rigorous psychometric methodologies. To

affirm its validity, the study concentrated on a specific facet: the evaluation of teacher competency traits. Fine-grained scoring guidelines were formulated, and state-of-the-art NLP techniques were used for text feature recognition and classification. The primary findings of this investigation can be summarized as follows: (1) Open-ended SJTs can establish precise scoring criteria grounded in crucial behavioral response elements; (2) Sentence-level text classification outperforms document-level classification, with CNNs exhibiting remarkable accuracy in response categorization; and (3) The scoring model consistently delivers robust performance and demonstrates a remarkable degree of alignment with human scoring, thereby hinting at its potential to partially supplant manual scoring procedures.

Keywords: machine learning, situational judgment tests, automated scoring, teacher competency, open-ended tests

Introduction

In the field of personnel assessment, Situational Judgment Tests (SJTs) have become widely popular due to their unique content and high face validity, frequently utilized for personnel selection and evaluation. Test items typically present a series of work-related scenarios, with options representing several typical behavioral responses, requiring test-takers to select the option that best matches their actual practice or to rank the options (Qi & Dai, 2003). SJTs serve as excellent tools for measuring competency, offering lower costs than interviews and greater vividness than self-report inventories, while demonstrating superior predictive validity for job performance compared to general cognitive ability tests and personality assessments (Burrus et al., 2012; McDaniel et al., 2007; McDaniel et al., 2011; Oostrom et al., 2012; Slaughter et al., 2014; Weekley & Ployhart, 2005).

Based on varying degrees of openness, SJT response formats can be broadly categorized into closed-response formats and open-ended formats. Closed-response formats represent the traditional multiple-choice (MC) approach, whereas open-ended formats, also known as constructed response (CR) formats, do not present options, allowing test-takers to respond freely. These primarily include written-constructed responses, audio-visual constructed responses, and situational interviews. Written-constructed formats require test-takers to articulate their approach in writing; audio-visual constructed formats typically present scenarios through multimedia, requiring verbal responses or performances that are recorded (Oostrom et al., 2010, 2011); situational interviews involve question-and-answer sessions between examiners and test-takers conducted face-to-face or online.

Closed-response formats currently dominate as the mainstream testing approach, facilitating standardized processing and rapid scoring. However, this format is susceptible to individual response attitudes, guessing, and test-taking strategies. Test-takers can glean cues from the options, and in high-stakes

contexts, may engage in socially desirable responding, making it difficult to effectively distinguish high-competency individuals (McDaniel et al., 2001; Robson et al., 2007). Furthermore, for test-takers, the options themselves impose additional cognitive load, requiring them to read through all options, parse their meanings, and make comparative judgments—a process in which extraneous variables such as cognitive ability may influence test outcomes (Lievens et al., 2015; Marentette et al., 2012).

Open-ended responding can address these issues to some extent. This format is not constrained by fixed answers, granting test-takers greater freedom of expression (Finch et al., 2018), promoting deeper understanding of thematic materials (Bacon, 2003; Rogers & Harley, 1999; Kastner & Stangla, 2011), and fostering higher motivation and more immersive responding (Arthur et al., 2002; Edwards & Arthur, 2007). Open-ended SJT items impose lower cognitive load and minimize guessing, demonstrating more desirable criterion-related validity (Funke & Schuler, 1998) and predictive validity (Arthur, 2002; Funke & Schuler, 1998; Lievens et al., 2019) compared to traditional multiple-choice formats, while more closely approximating real-life thinking and behavioral processes with higher ecological and face validity (Kjell et al., 2018).

Despite technological advances enabling increasing exploration of open-ended SJTs, current research remains in its nascent stages (Cucina et al., 2015). Researchers have examined written-constructed (Lievens et al., 2019) and audio-visual constructed formats (Oostrom et al., 2010, 2011), representing innovative attempts; however, the scoring phase still relies on manual methods. Manual scoring entails substantial time and labor costs (Edwards & Arthur, 2007; Downer et al., 2019; Iliev et al., 2015) and is susceptible to rater effects (Edwards & Arthur, 2007; Lievens et al., 2019). In Lievens et al.'s (2019) study, raters spent an average of approximately 35 minutes per test-taker, while Funke and Schuler (1998) employed three raters to ensure scoring quality. Consequently, in large-scale assessments where efficiency is paramount, such open-ended tests are often selected with caution. Scoring issues have become a significant impediment to the development of open-ended SJTs (Iliev et al., 2015), necessitating urgent resolution of automated scoring challenges.

Compared to manual scoring, automated scoring accommodates more diverse assessment tasks at lower cost while enabling immediate feedback. However, research on implementing automated scoring for open-ended SJTs remains scarce, with no established procedures or systematic research paradigms. Guo et al. (2021) utilized Natural Language Processing (NLP) techniques to analyze publicly available data from five open-ended SJTs, employing Doc2Vec to convert text into vectors and ridge regression to predict personality scores, yielding a modest average correlation coefficient of 0.28 ($r = 0.22$~0.38) without reporting the method's reliability or validity. Tavoosi (2022) designed a four-item open-ended SJT for Counterproductive Work Behavior (CWB) and employed N-gram methods for topic modeling to extract thematic words but did not implement actual scoring.

Although no explicit research paradigm exists, relevant studies offer methodological guidance. First, scoring criteria for open-ended SJTs can reference manual scoring standards. Manual scoring typically employs simple scoring rubrics requiring two or more raters; Lievens et al. (2019) utilized Behavioral Anchored Rating (BAR) scales for more concrete and objective criteria, a behavioral measurement tool for employee performance rating originally proposed by Smith and Kendall in 1963. Second, automated scoring algorithms can be categorized by text length: long-text formats such as Automated Essay Scoring (AES) and short-text formats such as Automatic Short-answer Grading (ASAG), with open-ended SJT automated scoring falling between these two categories. Third, automated scoring requires interpretability and validity verification. Psychometrics emphasizes scoring reliability, validity, and fairness; high model accuracy alone cannot sufficiently demonstrate machine scoring effectiveness. Evaluation metrics for machine scoring include correlation with human ratings, exact agreement rates, kappa coefficients, consistency of score distributions, and t-tests for related sample rating differences (Ramineni et al., 2012). Williamson et al. (2012) proposed a validity verification framework for machine scoring encompassing five aspects: interpretation, evaluation, extrapolation, generalization, and use of scoring results.

Among these considerations, automated scoring algorithms constitute the core of this study. AES and ASAG address different problem scenarios and evaluation emphases: AES focuses on assessing text conception, structure, writing style, grammar, and coherence, with high openness and text feature extraction as the core scoring component (Rudner & Liang, 2002; Yang et al., 2022). ASAG texts typically consist of several words or short phrases with reference answers, exhibiting lower openness as limited responses around standard answers, assessing specific knowledge points. Common methods include keyword matching (more keywords yield higher scores) or similarity algorithms (higher similarity to standard answers yields higher scores).

Open-ended SJT automated scoring represents a novel problem type, distinguished by: (1) Different degrees of openness—neither fully divergent (SJT responses can be categorized into limited categories) nor possessing standard answers. Individuals develop unique solutions under identical scenarios without explicit, knowledge-based "correct" answers (Whetzel & McDaniel, 2009). (2) Different scoring criteria—the focus of text mining lies in the relationship between natural language text and measured psychological traits, where different approaches under a given scenario represent varying ability levels and trait tendencies, with these differential tendencies constituting the assessment focus. Consequently, open-ended SJT automated scoring cannot directly reference existing algorithms: text style analysis cannot establish practical connections with measured psychological traits; keyword methods focus on surface semantic similarity and are unsuitable for semantically richer SJT responses; similarity algorithms are inappropriate as open-ended SJTs logically lack standard answers, and employing such methods would contravene the original intent of SJT item design.

Given that response texts contain different types of approaches, we can address scoring through semantic content analysis by transforming the scoring problem into a text classification task (Lubis et al., 2021; Ramesh & Sanampudi, 2022; Süzen et al., 2020). Automated Text Classification involves automatically assigning texts to predefined categories (Basu & Murthy, 2013). The text classification process primarily includes text preprocessing, feature extraction, model training, model evaluation, and model optimization/application. This supervised text classification process is illustrated in Figure 1 [Figure 1: see original paper], comprising two stages: first, model training on labeled training data; second, applying the trained model to test data for prediction and performance evaluation. In both stages, text data undergo identical preprocessing and feature extraction operations, such as stop-word removal and term frequency statistics, to obtain numerical text representations directly computable by machines. The trained classification model can be viewed as a mapping function from text representations to classification labels, obtained through specified machine learning algorithms to predict labels associated with classification texts.

Machine learning, particularly deep learning models, achieves favorable results in text classification tasks (Yang et al., 2022). Common machine learning classification algorithms include Support Vector Machines (SVM), k-Nearest Neighbor (KNN), Naive Bayes, and Decision Trees. In recent years, deep neural network-based text classification methods have achieved substantial breakthroughs, demonstrating more powerful performance. Deep learning methods employ pre-trained word vector models using deep neural networks such as Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) to implement text classification tasks. With sufficient corpora, these methods can exhibit excellent performance, achieving near-human levels in text scoring tasks and even demonstrating stronger stability than human raters.

In summary, open-ended situational judgment tests offer irreplaceable advantages, particularly suitable for scenarios requiring fine-grained individual profiling. These free-response texts contain rich emotional information and represent personality traits and behavioral tendencies. Mining text content enables more comprehensive psychological measurement and personalized evaluation. However, scoring presents certain difficulties: (1) Scoring standard development currently relies heavily on expert experience; (2) Automated scoring implementation is challenging, as evaluating free text is inherently difficult, and in psychological assessment contexts, computers' lack of understanding of response meanings further complicates automated scoring (Kastner & Stangla, 2011; Zhang et al., 2020); (3) Interpretation and validity verification of automated scoring remain understudied, with difficulties explaining the meaning of predictive scores output by scoring models and validity verification issues awaiting further research.

This study explores the application of open-ended SJTs in teacher competency assessment, targeting primary and secondary school teachers. Based on a psychometric framework, we developed an open-ended SJT, designed scoring cri-

teria incorporating typical behavioral responses, and implemented automated scoring using deep learning models. The automated scoring process comprises three stages: (1) Establishing scoring rules—based on manual coding, scoring rules are determined item-by-item according to key behaviors in each scenario, including behavioral response items and corresponding point values; (2) Automated text classification—modeling at both document and sentence levels, comparing multiple models through experiments, and selecting simple yet effective classification models to score all items; (3) Scoring performance verification—validating scoring effectiveness from multiple perspectives including model performance and machine-scoring reliability and validity. The specific process is illustrated in Figure 2 [Figure 2: see original paper].

Research expectations: (1) The independently developed open-ended teacher competency SJT will demonstrate satisfactory reliability and validity, effectively distinguishing teacher competency levels; (2) Deep learning-based text classification models can be applied to such subjective scoring tasks without standard answers, achieving high machine-scoring accuracy; (3) Machine scoring will exhibit good reliability and validity, with strong positive correlations between human and machine ratings.

2. Methodology

2.1 Participants

A total of 627 primary and secondary school teachers from Shenzhen participated in the testing (age: 26-40 years, M = 31.52 years, SD = 2.2), including 463 females and 164 males. Teachers of Chinese, mathematics, and English accounted for 42.9%, with teachers of other subjects comprising 57.1%.

2.2.1 Open-ended Teacher Competency Situational Judgment Test

The test development process proceeded as follows. First, test dimensions were established using a classical procedure (Xu, 2004) to construct competency characteristics for primary and secondary school teachers. Behavioral Event Interviews (BEI) were conducted with 12 frontline teachers from 8 primary and secondary schools in Beijing, including 7 females and 5 males, with 6 being key teachers. Interviewees were guided to recall their most successful and most regrettable career events, with each interview lasting 2-3 hours. After transcribing and organizing interview recordings and texts, frequently occurring key competency characteristics were categorized and summarized. The final competency model comprised 4 primary dimensions and 10 secondary dimensions: (1) Student-oriented: caring for students and developing others; (2) Problem-solving: dynamic decision-making and flexible adaptation; (3) Emotional intelligence: understanding others, emotional control, and interpersonal communication; (4) Achievement motivation: responsibility, challenging difficulties, and perseverance.

Item development was based on literature and interviews to identify five typical problem scenarios in teaching: student management, classroom instruction, colleague relationships, student counseling, and home-school communication. According to the four primary dimensions, 54 representative problem scenarios were selected and developed into stems and options, uniformly employing the instruction: "In such a situation, what would you do?"

Expert evaluation and item revision involved distributing expert evaluation questionnaires to 54 experienced primary school teachers in Henan Province, with 88.24% having over 10 years of teaching experience, yielding 34 valid questionnaires. This version consisted of single-choice items with 4 options. In addition to completing the test, participants completed an evaluation questionnaire assessing scenario authenticity (5-point scale), evaluated options by identifying actual, optimal, and worst practices, and provided supplementary approaches and revision suggestions. Statistical analysis revealed a mean scenario authenticity rating of 3.61 (out of 5). Analysis of option distributions revealed a pronounced tendency toward dominant responding. Based on expert feedback, items were revised, resulting in a final open-ended SJT comprising 20 items across 4 dimensions: Student-oriented (items 1, 8, 9, 10, 12, 16, 20), Problem-solving (items 3, 4, 6, 7, 17, 18), Emotional intelligence (items 2, 5, 11, 19), and Achievement motivation (items 13, 14, 15).

2.2.2 Criterion Measures

Job Satisfaction Questionnaire. The Teacher Job Satisfaction Scale developed by Feng (1996) was employed, comprising 26 items across 5 dimensions: self-actualization, workload, salary, leadership relationships, and colleague relationships. Reliability and validity were examined using the current dataset, yielding an overall Cronbach's alpha of 0.89 (N = 627). Alpha coefficients for the five dimensions were: self-actualization 0.84, workload 0.76, salary 0.77, leadership relationships 0.79, and colleague relationships 0.73. Confirmatory factor analysis results were: ^{2} = 1055.595, df = 289, ^{2}/df = 3.65, RMSEA = 0.065, CFI = 0.868, TLI = 0.851, SRMR = 0.063.

General and Subject-Specific Teaching Philosophy Questionnaires. The General Teaching Philosophy questionnaire contained 12 items. Confirmatory factor analysis led to removal of 2 items with factor loadings below 0.3 (items 2, 12), retaining 10 items. Analysis yielded an overall Cronbach's alpha of 0.88 (N = 627) with good model fit (^{2} = 131.363, df = 35, ^{2}/df = 3.75, RMSEA = 0.066, CFI = 0.964, TLI = 0.954, SRMR = 0.029). Subject-specific teaching philosophy was assessed separately for Chinese, mathematics, and English, with alpha coefficients of 0.93 (n = 99), 0.68 (n = 86), and 0.78 (n = 84), respectively.

Comprehensive Teaching Performance Assessment Materials. Complete materials were submitted by 181 participants and rated by 6 teaching experts on a 3-point scale per dimension. Assessment materials covered pre-

, during-, and post-instruction work, specifically including: (1) Instructional design—teachers provided a lesson plan according to uniform requirements, evaluated across 6 aspects: instructional rationale, objectives, key points, difficulties, methods, and process; (2) Instructional video—a complete 30+ minute classroom recording scored across 4 dimensions (classroom management, instructional content, thinking cultivation, emotional attention) using a classroom observation scale (Ling, 2020). The scale's Cronbach's alpha was 0.83, with dimension alphas of 0.67, 0.65, 0.41, and 0.69, respectively. Confirmatory factor analysis results were: $^{2} = 150.12$, df = 82, $^{2}/\text{df} = 1.83$, RMSEA = 0.075, CFI = 0.897, TLI = 0.868, SRMR = 0.060. (3) Student assignments—teachers submitted 9 representative student assignments (3 each of excellent, good, and poor quality) for expert evaluation of assignment content design, evaluation criteria design, and student work analysis.

2.3 Data Analysis

SPSS 26.0 and Mplus 8.3 were used for test quality analysis, Nvivo 11 software for manual coding, and Python 3.8 for data training and prediction.

2.4 Scoring Process

2.4.1 Problem Definition The foremost consideration in text scoring is the establishment of scoring criteria. Open-ended SJT response texts are characterized by: containing several approaches within a single answer, including problem-solving steps, logic, and sequences, without possessing a single, definitive answer. The core of scoring is not whether the approach in the scenario is correct, but the degree of alignment between typical behavioral response patterns in the text and the teacher competency model. This study does not establish answer templates but adopts a Behavioral Anchored Rating approach, focusing on key behaviors triggered by specific stimuli in scenarios. Coders categorize all responses in the text and further cluster categories into typical behavioral response sets, assigning point values to each response item. Since key behaviors differ across scenarios, scoring rules must be established separately for each item.

2.4.2 Manual Coding Response texts from 300 participants were selected for manual coding, excluding 10 participants with response times under 1000 seconds or obvious non-serious responses (repetitive or irrelevant text), retaining 290 texts. Four psychology graduate students served as coders, receiving half-day unified training covering assessment dimensions, coding standards, software operation, and principles for handling disputed items, with items randomly assigned.

The coding process comprised two components: First, identifying behavioral response items—specifically, two coders independently read through texts for each item, identifying all behavioral response items, then collaboratively revised and merged them, clustering response items to establish typical behavioral response

categories (10-30 categories, typically around a dozen). Second, manual coding annotation (labeling)—one coder annotated sentence-by-sentence in Nvivo software, while the other verified coding results, with opportunities to raise disagreements and further refine coding rules. Upon completion of coding for each item, results were exported and organized into sentence annotation datasets by participant ID.

2.4.3 Scoring Rule Development Building upon the behavioral response items obtained in the previous step, point values were assigned to each response item. Based on the degree of match between responses and competency characteristics, higher point values were assigned to response items more closely aligned with competency traits, while simultaneously considering differences in thinking level and ability reflected by behavioral richness, specificity, comprehensiveness, and logicality. A 3-point scale (0-3) was employed: 1 = poor, 3 = excellent, 0 = off-topic or invalid response. Point assignment for each item was determined through discussion between two raters until consensus was reached.

2.4.4 Score Composition Based on behavioral response items obtained for each participant ID during manual coding and corresponding point values from the scoring rules, each behavioral response item was converted to a score item-by-item. A single response typically contained multiple behavioral response items, with raw total scores for each item synthesized by summing these scores and converted to grade scores based on percentiles—the top 27% received grade scores of 3, the bottom 27% received 1. Additionally, dimension scores and total test scores were calculated.

2.5 Automated Scoring Implementation

2.5.1 Dataset and Evaluation Metrics Annotated texts from participants ID 1-300 were selected as the dataset, comprising 20 items and 6,000 responses. For each item, texts were divided into training and test sets at a 2:1 ratio across 300 participants. In machine learning, classification tasks are typically evaluated using accuracy, precision, recall, and F1-score. Below is an illustration of these four metrics using binary classification as an example. Assuming binary classification includes positive and negative classes, Table 1 presents the confusion matrix for binary classification, with matrix elements defined as: (1) TP (True Positive): number of samples actually positive and predicted positive; (2) TN (True Negative): number of samples actually negative and predicted negative; (3) FP (False Positive): number of samples actually negative but predicted positive; (4) FN (False Negative): number of samples actually positive but predicted negative.

Table 1 Confusion Matrix for Binary Classification

	Predicted Positive	Predicted Negative
Actual Positive	TP (True Positive)	FN (False Negative)
Actual Negative	FP (False Positive)	TN (True Negative)

Accuracy reflects model performance across all samples, equal to the number of correctly classified samples divided by total samples—the sum of diagonal elements divided by all matrix elements: Accuracy = (TP + TN) / (TP + FN + FP + TN). Precision, recall, and F1-score are calculated separately for each class. For the positive class in binary classification, precision equals the number of positive samples predicted as positive divided by all samples predicted as positive: P = TP / (TP + FP). Recall equals the number of positive samples predicted as positive divided by actual positive samples: R = TP / (TP + FN). F1-score is the harmonic mean of precision and recall: F1 = 2PR / (P + R). The text classification in this study primarily involves multi-class tasks; evaluation metrics are calculated by first computing P, R, and F1 for each class separately, then calculating weighted averages based on sample sizes per class to obtain final precision, recall, and F1-scores.

2.5.2 Document-Level Multi-Label Text Classification Traditional text classification tasks typically involve single-label learning, where each text belongs to only one mutually exclusive category label marked with 0 or 1. However, many samples simultaneously belong to multiple labels from a label set. Schapire (1999) proposed multi-label learning, assigning each instance the most relevant subset of class labels from the label set. Based on the characteristic that a single response text contains multiple categories of behavioral response items, we first attempted multi-label classification at the document level. Using the first item as an experiment, deep learning algorithms were employed for classification modeling to achieve automated mapping from response texts to label systems. In implementation, data preprocessing first removed stop words. Input texts were converted to numerical matrix form through Jieba segmentation and Word2vec pre-trained word vectors, then connected to neural network layers with trainable parameters, fully connected layers, and SoftMax layers, ultimately outputting probabilities of the text belonging to various labels. Multiple deep learning methods were applied in the neural network layer, including Convolutional Neural Networks (CNN) (Kim, 2014), Recurrent Neural Networks (RNN) (Zhao et al., 2019), Recurrent Convolutional Neural Networks (R-CNN) (Lai et al., 2015), and RNN with Attention (Pang et al., 2021). CNN primarily captures local deep text features for various labels through convolutional kernel parameters; RNN captures global deep text features through recurrent unit structures; R-CNN combines both advantages by serially connecting RNN and CNN; Attention optimizes deep text representation by calculating weights for each word in the text through neural networks, typically used in series with RNN.

2.5.3 Sentence-Level Text Multi-Classification Multi-label classification tasks output multiple labels for entire response texts at the document level. If documents are segmented, individual labels can be output for each sentence at the sentence level. During manual coding, sentence-by-sentence annotation datasets were already obtained.

Four items were randomly selected. Data preprocessing first segmented sentences using punctuation marks ("。.!?; ;") and enumerations ("— (—) 1(1) "), removed stop words, and converted texts to numerical matrices through Jieba segmentation and Word2vec pre-trained word vectors. Four deep learning models were trained: Convolutional Neural Networks (CNN), Long Short-Term Memory Networks (LSTM) (Hochreiter et al., 1997), CNN-LSTM (C-LSTM), and LSTM with Attention (LSTM+attention), conducting both score prediction and behavioral response item prediction. LSTM effectively addresses gradient vanishing and explosion problems as a structural variant of RNN; C-LSTM combines CNN and LSTM (Zhou et al., 2015), capturing both local sentence features and temporal sentence semantics across full texts. Models learned sentence sets corresponding to each response item or score in the annotation set to identify deep semantic relationships between texts, thereby completing model training. Each sentence produced two prediction outcomes: score prediction (outputting sentence scores from 0-3) and label prediction (behavioral response items), enabling more detailed evaluation of test-takers' thoughts and abilities.

3. Results

3.1 Scoring Rules

In the original response dataset, each item's text ranged from 100-300 characters, totaling 1,353,365 characters across 20 items. The first 300 responses were coded, comprising 647,322 characters, with individual items containing 724-1,453 sentences and a total of 19,368 annotated sentences. Inter-coder reliability was examined using the first item, yielding r = 0.84 and quadratic weighted kappa = 0.78 between two raters.

Scoring rules for each item were generated following manual coding, primarily comprising two components: typical behavioral response items in the scenario and their corresponding point values. Each response item received a unique identifier, resulting in 20 scoring rules.

3.2 Test Quality Analysis

Multiple reliability indices were used to examine multidimensional test reliability (Gu & Wen, 2017). Under a bi-factor structure treating competency as a global factor and four dimensions as local factors, the homogeneity coefficient (HC) and total composite reliability were 0.88 and 0.96, respectively. The overall Cronbach's alpha was 0.91, with dimension alphas of: Student-oriented 0.79, Problem-solving 0.76, Emotional intelligence 0.66, and Achievement motivation 0.60.

To examine structural validity, four confirmatory factor analysis models were specified and compared: M1 as a single-factor model with all items loading on one factor; M2 as a four-factor model; M3 as a bi-factor model (BFM) with all items additionally loading on a global factor orthogonal to local factors; and M4 as a bi-factor model with orthogonal global and local factors but correlated local factors. Results in Table 2 indicate M4 was clearly superior, thus selected as the best model. The test exhibited a clear bi-factor structure with one competency global factor and four dimensions.

Table 2 Confirmatory Factor Analysis of Teacher Competency Situational Judgment Test (n = 290)

Model	^{2}/df	SRMR	RMSEA

Criterion-related validity was examined using job satisfaction, teaching philosophy, and teaching ability as criteria. Results in Table 3 show competency total scores significantly correlated with job satisfaction ($r_1 = 0.20$, $p = 0.001$), general teaching philosophy ($r_2 = 0.21$, $p < 0.001$), subject-specific teaching philosophy ($r_3 = 0.22$, $p < 0.001$), instructional design ($r_4 = 0.26$, $p < 0.001$), classroom evaluation ($r_5 = 0.20$, $p = 0.007$), and student assignments ($r_6 = 0.22$, $p = 0.003$).

Table 3 Correlation Analysis Between Teacher Competency Total/Dimensions and Criterion Variables

Variable	M ± SD	1	2	3	4	5	6	7	8
Competency Total	37.77 ± 8.17	0.89***	0.87***	0.78***	0.75***	0.20**	0.21***	0.22***	-
Student-oriented	1.89 ± 0.46	0.87***	0.68***	0.58***	0.55***	0.22***	0.18**	0.21**	0.68***
Problem-solving	1.87 ± 0.48	0.88***	0.68***	0.53***	0.52***	0.24**	0.21**	0.18*	0.58***
Emotional intelligence	1.91 ± 0.50	0.78***	0.58***	0.55***	0.14*	0.15*	0.18**	0.54***	0.55***
Achievement motivation	1.89 ± 0.54	0.75***	0.55***	0.52***	0.54***	0.27***	0.17*	0.22**	0.54***
Instructional design	3.88 ± 0.29	0.20**	0.26***	0.24**	0.14*	0.15*	0.20***	0.13*	0.15*
Classroom evaluation	4.58 ± 0.41	0.21***	0.20**	0.21**	0.18**	0.20***	0.13*	0.18**	0.20**
Student assignments	3.58 ± 0.49	0.22***	0.22**	0.18*	0.16*	0.18**	0.45***	0.33***	0.50***

Note: *$p < 0.05$*, ***$p < 0.01$***, $p < 0.001$; Individual SJT items used 3-point

scoring, total test score range = 0-60; Job satisfaction and teaching philosophy questionnaires used 5-point scales; Expert ratings for instructional design, classroom video, and student assignments used 3-point scales per dimension.

3.3 Automated Scoring Performance

3.3.1 Document-Level Multi-Label Text Classification Using multi-label annotation methods, entire responses were output with multiple labels. Experimental results in Table 4 show all models performed inadequately on the test set, with accuracies of 46%-55%. Researchers speculated that this resulted from limited sample sizes and excessive classification categories (approximately 20 categories per item), with most labels being tail labels with only a few annotations.

Table 4 Comparison of Document-Level Multi-Label Text Classification Model Performance

Model	Accuracy	Precision	Recall	F1
CNN	-	-	-	-
RNN	-	-	-	-
R-CNN	-	-	-	-
RNN+Attention	-	-	-	-

3.3.2 Sentence-Level Text Multi-Classification Response texts were segmented into sentence units for model training on four randomly selected items. Experimental results indicated: (1) For score prediction tasks, on Item 20, the four algorithms showed small differences in accuracy and F1-score, with CNN achieving highest precision and C-LSTM highest recall; on Items 6 and 7, four metrics showed minor differences, with C-LSTM slightly better on Item 6 and LSTM slightly better on Item 7; on Item 3, CNN clearly outperformed other models. (2) For response item prediction tasks, on Item 20, the four algorithms showed small differences in accuracy and F1-score, with CNN showing higher precision and LSTM higher recall; on Item 6, F1-scores and recall showed minor differences, with CNN and LSTM showing higher accuracy and CNN highest precision; on Items 7 and 3, CNN achieved best performance across all four metrics. Overall, CNN performed best, with predicted score accuracies of 79%-92% and predicted response item accuracies of 75%-80% across four items, as shown in Figure 3 [Figure 3: see original paper].

(a) Score Prediction
(b) Response Item Prediction

Figure 3 Comparison of Four Models on Response Item and Score Prediction Tasks Across Four Items

Note: Acc = Accuracy; F1 = F1-score; Pre = Precision; Rec = Recall; same below.

3.3.3 Overall Performance Sentence-level accuracy surpassed document-level accuracy; therefore, the sentence-level text multi-classification approach was adopted, selecting the best-performing CNN model for automated scoring of all items. Results in Figure 4 [Figure 4: see original paper] show computer-predicted score accuracies of 70%-88% across 20 items, representing good performance. Predicted behavioral response item accuracies ranged from 58%-81%, which remains respectable given limited training corpora, rich semantic characteristics, and numerous classification categories (10-20+ categories).

(a) Score Prediction
(b) Response Item Prediction

Figure 4 CNN Performance Across 20 Items

3.4 Validity Verification of Automated Scoring

Data from the first 200 participants in the annotated set served as the training set, with data from the remaining 100 participants as the test set. Among the 100 participants' machine-scored results, 6 with incomplete data or excessively short response times were removed, leaving 94 participants with 1,880 responses for comparative human-machine scoring analysis to examine machine-scoring reliability and validity.

3.4.1 Human-Machine Scoring Agreement Score distributions for human and machine ratings were similar. Human-rated total scores (36.36 $\pm$ 7.99) showed kurtosis of -0.592 and skewness of 0.175, while machine-rated total scores (37.23 $\pm$ 7.83) showed kurtosis of -0.345 and skewness of 0.151, as shown in Figure 5 [Figure 5: see original paper].

Figure 5 Frequency Distribution of Total Scores for Human and Machine Ratings

Correlation analysis revealed that human-rated total scores (36.36 $\pm$ 7.99) and machine-rated total scores (37.23 $\pm$ 7.83) were highly positively correlated ($r = 0.95$, $p < 0.001$). All four dimensions showed high positive correlations between human (1.81 $\pm$ 0.45, 1.82 $\pm$ 0.49, 1.87 $\pm$ 0.49, 1.78 $\pm$ 0.54) and machine ratings (1.89 $\pm$ 0.44, 1.84 $\pm$ 0.47, 1.87 $\pm$ 0.47, 1.82 $\pm$ 0.53): Student-oriented $r = 0.91$, Problem-solving $r = 0.90$, Emotional intelligence $r = 0.81$, Achievement motivation $r = 0.89$ (all $p < 0.001$), meeting requirements for large-scale assessment use. Human-machine correlations for 20 items ranged from 0.48 to 0.90 (all $p < 0.001$).

Quadratic Weighted Kappa (QWK) coefficients were employed as evaluation criteria. Williamson et al. (2012) suggest automated scoring QWK should be at least 0.7 for high-stakes testing. This study's human-machine QWK—total score (0.82) and dimensions (Student-oriented 0.89, Problem-solving 0.90, Emotional intelligence 0.81, Achievement motivation 0.89)—all met criteria for high-stakes test use.

3.4.2 Machine Scoring Reliability and Validity Internal consistency reliability was measured using Cronbach's alpha calculated from machine-rated results: overall test $\alpha = 0.87$, dimension αs: Student-oriented 0.66, Problem-solving 0.73, Emotional intelligence 0.55, Achievement motivation 0.55.

Confirmatory factor analysis (n = 94) yielded: $^2 = 210.896$, df = 164, $^2/df = 3.75$, RMSEA = 0.055, CFI = 0.884, TLI = 0.866, SRMR = 0.029, with factor loadings ranging from 0.412-0.659, indicating adequate but lower structural validity than human scoring. Correlations between machine-rated total and dimension scores are shown in Table 5 , with high correlations between dimensions and total score and moderate inter-dimensional correlations. Criterion-related validity showed competency total scores significantly correlated with general teaching philosophy (r = 0.22, p = 0.036).

Table 5 Descriptive Statistics and Correlation Analysis of Machine Ratings (n = 94)

Variable	M ± SD	1	2	3	4	5	6	7	8
1. Machine-rated total	37.23 ± 7.83	-	-	-	-	-	-	-	-
2. Student-oriented	1.89 ± 0.44	0.90***	-	-	-	-	-	-	-
3. Problem-solving	1.84 ± 0.47	0.88***	0.72***	-	-	-	-	-	-
4. Emotional intelligence	1.87 ± 0.47	0.80***	0.63***	0.58***	-	-	-	-	-
5. Achievement motivation	1.82 ± 0.53	0.72***	0.53***	0.58***	0.54***	-	-	-	-
6. General teaching philosophy	3.85 ± 0.27	0.22*	0.27*	0.24*	0.13*	0.20**	-	-	-
7. Subject teaching philosophy	4.28 ± 0.39	0.51***	0.63***	0.32**	0.44***	0.18*	0.69***	-	-
8. Job satisfaction	3.65 ± 0.58	0.54***	0.58***	0.53***	0.39***	0.61***	0.43***	0.49***	-

4. Discussion

This study attempted to explore a psychometric theory-based paradigm for automated scoring of open-ended situational judgment tests. To validate automated scoring effectiveness, we focused on the specific research problem of teacher competency assessment, developed an open-ended SJT, established fine-grained scoring rules, employed NLP technology for text feature recognition and classification, and explored automated scoring methods using various deep

learning models (CNN, RNN, R-CNN, RNN+Attention, LSTM, C-LSTM, LSTM+Attention) at both document and sentence levels. Results indicate that sentence-level classification outperformed document-level classification, with CNN demonstrating superior performance. Model-predicted score accuracies reached 70%-88%, and predicted response item accuracies reached 58%-81%, indicating good model performance capable of accurate automated scoring. Specific issues are discussed below.

4.1 Scoring Standard Design

Problem definition must precede scoring standard establishment, determining scoring strategies based on test content/type and response text characteristics. For instance, the presence or absence of standard answers determines scoring logic and algorithm design, while response length and semantic richness determine whether manual coding is needed and influence scoring strategies. Scoring rules should maximally reflect individual trait-level information, with two key considerations: (1) Reasonable categorization—behavioral response items should be comprehensive, specific, and representative, covering all possible types. Categories must sufficiently reflect differences while avoiding excessive granularity that introduces randomness. Detailed, specific response items better reflect differences and discriminability, but excessive categories reduce prediction accuracy; conversely, fewer behavioral response items improve prediction accuracy but reduce discriminability. (2) Reasonable point assignment—score levels should reflect test-taker proficiency, with point assignment for each behavioral response item being a difficult process requiring repeated deliberation and comprehensive consideration.

Furthermore, test quality directly impacts scoring effectiveness. Test development and automated scoring are not independent processes. The automated scoring method explored in this study does not prioritize model complexity or perfection but rather designing a feasible open-ended SJT automated scoring approach—establishing reasonable scoring rules and selecting appropriate scoring models to gradually improve accuracy. Open-ended responding does not imply arbitrary test development; rather, items should be developed according to standardized procedures based on qualified, reliable, and valid tests. Test developers must possess deep understanding of assessed dimensions, grasping trait connotations and behavioral manifestations through extensive research and interviews to design effective scoring rules. Additionally, item wording should avoid ambiguity and excessive extraneous or distracting information that could compromise test quality.

4.2 Automated Scoring Process

Multiple methods and models were employed for comparative experiments to select optimal models. Based on specific task input-output formats, various modeling approaches exist for automated scoring, requiring practical experimentation to identify simpler, more effective methods. In this study, input

comprised test-taker response texts, output comprised multiple response items or multi-level scores—directly corresponding to multi-label classification tasks in machine learning. Therefore, document-level multi-label text classification was first attempted. This approach did not incorporate sentence-level annotation information and would be preferable if adequate performance were achieved. However, in practice, multi-label classification yielded suboptimal results, with the process illustrated using only the first item. Sentence-level automated scoring achieved more effective results and was consequently adopted.

Different deep learning models possess unique advantages and limitations in text classification tasks, producing varying impacts on automated scoring performance. For example, CNN primarily captures local features in text such as phrases and collocations, potentially performing poorly on tasks requiring long-range dependency consideration due to ineffective handling of global information in long text sequences. RNN and its variants (LSTM, GRU) capture contextual information when processing sequential data, suitable for tasks with strong long-term dependencies in text. However, traditional RNN struggles with long sequences; although LSTM and GRU mitigate these issues to some extent, they remain constrained by text length and perform poorly on some text analysis tasks. Attention mechanisms enable models to focus on key text parts, helping capture important information, but early attention mechanisms were typically bound to RNN models and subject to their limitations. In this study, since sentence-level response item classification tasks are typically associated with specific phrases and collocations, CNN's optimal performance is understandable. Across diverse research tasks, different deep learning models exhibit distinct characteristics in automated scoring, with model selection significantly impacting performance. Selecting appropriate models based on task-specific requirements and model advantages/limitations helps improve automated scoring accuracy. Furthermore, with the introduction of pre-trained language models and large language models (e.g., ChatGPT), automated scoring models have richer options, but given scenario specificity, rigorous performance evaluation and validity verification remain necessary to determine model usability.

Automated scoring effectiveness is also influenced by multiple human factors. Data preprocessing requires careful attention to sentence segmentation methods. Sentence segmentation quality directly impacts scores; for machines, distinguishing different semantic units in a text is challenging, and more non-standard punctuation usage in datasets leads to poorer segmentation quality. Therefore, adding a validation step for segmented datasets after machine segmentation helps achieve better subsequent scoring results. In broader test types, appropriate segmentation markers should be selected based on text length and semantic complexity. Additionally, multiple approaches should ensure manual coding quality and optimize scoring rule establishment.

4.3 Validity and Interpretability of Automated Scoring

Human and machine scoring exhibit different characteristics in psychological testing use. Wang and Peng (2019) compared human-machine scoring features, finding humans performed better in identifying off-topic responses, memorized templates, semantic judgment, and recognizing response sequencing and logical order, while machine scoring showed less central tendency, stronger overall response grasp, and better ability to identify anomalous responses. Whether machine scoring results can assist or replace human scoring requires attention not only to model evaluation metrics such as prediction accuracy but also to scoring reliability and validity, particularly validity verification. In this study, human-machine r and QWK exceeded 0.8 for total and four dimension scores, with machine scoring demonstrating stronger stability than human scoring on some items (e.g., Item 1 human-machine agreement r = 0.88 exceeded inter-rater agreement r = 0.78). Therefore, the automated scoring system is effective and can replace at least one human rater in scoring processes, enabling human-machine combined scoring or fully automated scoring.

Validity research is viewed as a process of making plausible interpretations of test scores (Xie, 2013), with interpretability of automated scoring representing a particularly challenging research problem. Machine learning processes typically establish black-box models with weak interpretability, struggling to meet psychological tests' descriptive requirements for assessment elements. Scoring based solely on data itself and distances between text representations is insufficient. This study incorporated expert knowledge in model construction, which is key to converting the machine scoring process into a "white-box" model. The invisible scoring process is transformed into first classifying text into behavioral response items corresponding to scoring rules, yielding not only scores but also test-takers' behavioral response items. Based on these behavioral items, further mining of test-takers' behavioral patterns, thinking styles, and personality characteristics is possible. This behavior-focused scoring enables more detailed characterization of behavioral differences, representing a finer-grained scoring model with greater interpretability.

4.4 Practical Implications

This research holds broad application prospects and practical significance: First, it explores open-ended response formats for SJTs, reducing test-taker guessing and faking behaviors in multiple-choice SJTs and enabling more personalized individual assessment. Second, it explores automated scoring technology for Chinese subjective items without standard answers, establishing a complete paradigm: open-ended test development → classification and aggregation → coding annotation → expert scoring → automated scoring → effectiveness verification. Moving beyond simple semantic and similarity computation, it emphasizes the correspondence between text and measured psychological traits, enhancing automated scoring interpretability through manual coding and fine-grained scoring rules. Third, the automated scoring model achieves high accu-

racy and good performance, providing an efficient, time- and labor-saving, accurate, and reliable assessment tool in practice. Fourth, it expands open-ended SJT applications and provides reference and guidance for automated scoring of other open-ended item types, facilitating broader application of open-ended formats in testing contexts.

4.5 Limitations and Future Directions

This study has several limitations: First, sample representativeness—participants were primary and secondary school teachers from Shenzhen, representing young teachers in well-resourced areas with strong homogeneity, not generalizable to broader teacher populations. Second, limited annotation quantity—scoring precision is constrained by annotated sample size; due to time and labor constraints, each item contained approximately 1,000 annotated sentences with uneven label distribution, affecting machine learning effectiveness. Third, behavioral response items in scoring rules could benefit from further classification refinement and adjustment. Fourth, criterion selection—more appropriate criteria should be selected to validate scoring validity through multiple sources of evidence.

Future research will consider expanding the item bank, continuously updating problem scenarios emerging in contemporary teaching contexts while enhancing item specificity. Additionally, AI-assisted automatic coding methods will be explored to improve efficiency in response item categorization. In scoring algorithms, few-shot learning methods will be attempted to further improve machine-scoring accuracy. Furthermore, for richer open-ended constructed response formats such as speech and body movements, technical approaches from AI interview systems (Lee & Kim, 2021) can be referenced to explore broader application spaces for open-ended SJTs.

Under the conditions of this study, the following conclusions were drawn: (1) Open-ended situational judgment tests can establish scoring rules based on key behavioral response items, with automated scoring steps including: behavioral response item classification and aggregation → coding annotation → expert scoring → automated scoring → effectiveness verification; (2) Scoring algorithms can be designed at both document and sentence levels, with sentence-level text classification outperforming document-level classification in this study, where convolutional neural networks achieved higher classification accuracy by better capturing lexical features of key behavioral response items; (3) The developed scoring model demonstrated stable performance, high consistency between machine and human scoring, and good reliability and validity, enabling partial replacement of human raters in practical scoring tasks.

References

Arthur, W., Edwards, B. D., & Barrett, G. V. (2002). Multiple-choice and constructed response tests of ability: Race-based subgroup performance differences

on alternative paper-and-pencil test formats. *Personnel Psychology, 55*(4), 985–1008.

Bacon, D. R. (2003). Assessing learning outcomes: A comparison of multiple-choice and short-answer questions in a marketing context. *Journal of Marketing Education, 25*(1), 31–36.

Basu, T., & Murthy, C. A. (2013, December). Effective text classification by a supervised feature selection approach. *IEEE 12th International Conference on Data Mining Workshops (ICDM)*, 918–925, Brussels, Belgium.

Burrus, J., Betancourt, A., Holtzman, S., Minsky, J., MacCann, C., & Roberts, R. D. (2012). Emotional intelligence relates to well-being: Evidence from the situational judgment test of emotional management. *Applied Psychology: Health and Well-Being, 4*(2), 151–166.

Burrows, S., Gurevych, I., & Stein, B. (2015). The eras and trends of automatic short answer grading. *International Journal of Artificial Intelligence in Education, 25*, 60–117.

Cucina, J. M., Su, C., Busciglio, H. H., Thomas, P. H., & Peyton, S. T. (2015). Video-based testing: A high-fidelity job simulation that demonstrates reliability, validity, and utility. *International Journal of Selection and Assessment, 23*(3), 197–209.

Downer, K., Wells, C., & Crichton, C. (2019). All work and no play: A text analysis. *International Journal of Market Research, 61*(3), 236–251.

Edwards, B. D., & Arthur, W., Jr. (2007). An examination of factors contributing to a reduction in subgroup differences on a constructed-response paper-and-pencil test of scholastic achievement. *Journal of Applied Psychology, 92*(3), 794–801.

Finch, W. H., Finch, M. E. H., McIntosh, C. E., & Braun, C. (2018). The use of topic modeling with latent dirichlet analysis with open-ended survey items. *Translational Issues in Psychological Science, 4*(4), 403–424.

Funke, U., & Schuler, H. (1998). Validity of stimulus and response components in a video test of social competence. *International Journal of Selection and Assessment, 6*(2), 115–123.

Gu, H. L., & Wen, Z. L. (2017). Reporting and interpreting multidimensional test scores: A bi-factor perspective. *Psychological Development and Education, 33*(4), 504–512.

Guo, F., Gallagher, C. M., Sun, T., Tavoosi, S., & Min, H. (2021). Smarter people analytics with organizational text data: Demonstrations using classic and advanced NLP models. *Human Resource Management Journal*. Advance Online Publication.

Iliev, R., Deghani, M., & Sagi, E. (2015). Automated text analysis in psychology: methods, applications, and future developments. *Language and Cognition,*

7(2), 265–290.

Kastner, M., & Stangla, B. (2011). Multiple choice and constructed response tests: Do test format and scoring matter? *Procedia-Social and Behavioral Sciences, 12*, 263–273.

Kim, Y. (2014). Convolutional neural networks for sentence classification. *Proceedings of the 19th Conference on Empirical Methods in Natural Language Processing*, 1746–1751.

Kjell, O. E., Kjell, K., Garcia, D., & Sikstrom, S. (2018). Semantic measures: Using natural language processing to measure, differentiate, and describe psychological constructs. *Psychological Methods, 24*(1), 92–115.

Lai, S., Xu, L., Liu, K., & Zhao, J. (2015). Recurrent convolutional neural networks for text classification. *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, 2267–2273.

Lee, B. C., & Kim, B. Y. (2021). Development of an AI-based interview system for remote hiring. *International Journal of Advanced Research in Engineering and Technology, 12*(3), 654–663.

Lievens, F., De Corte, W., & Westerveld, L. (2015). Understanding the building blocks of selection procedures: Effects of response fidelity on performance and validity. *Journal of Management, 41*(6), 1604–1627.

Lievens, F., Sackett, P. R., Dahlke, J. A., Oostrom, J. K., & De Soete, B. (2019). Constructed response formats and their effects on minority-majority differences and validity. *Journal of Applied Psychology, 104*(5), 715–726.

Ling, C. (2020). *Development of Classroom Observation Scale to Promote the Professional Development of New Teachers* (Unpublished master's thesis). Beijing Normal University.

Lubis, F. F., Mutaqin, Putri, A., Waskita, D., Sulistyaningtyas, T., Arman, A. A., & Rosmansyah, Y. (2021). Automated short-answer grading using semantic similarity based on word embedding. *International Journal of Technology, 12*(3), 571–581.

Marentette, B. J., Meyers, L. S., Hurtz, G. M., & Kuang, D. C. (2012). Order effects on situational judgment test items: A case of construct-irrelevant difficulty. *International Journal of Selection and Assessment, 20*(3), 287–294.

McDaniel, M. A., Hartman, N. S., Whetzel, D. L., & Grubb, W. L. (2007). Situational judgment tests, response instructions, and validity: A meta-analysis. *Personnel Psychology, 60*(1), 63–91.

McDaniel, M. A., Morgeson, F. P., Finnegan, E. B., Campion, M. A., & Braverman, E. P. (2001). Use of situational judgment tests to predict job performance: A clarification of the literature. *Journal of Applied Psychology, 86*(4), 730–740.

McDaniel, M. A., Psotka, J., Legree, P. J., Yost, A. P., & Weekley, J. A. (2011). Toward an understanding of situational judgment item validity and group differences. *Journal of Applied Psychology, 96*(2), 327–336.

Oostrom, J. K., Born, M. P., Serlie, A. W., & Van der Molen, H. T. (2010). Webcam testing: Validation of an innovative open-ended multimedia test. *European Journal of Work and Organizational Psychology, 19*(5), 532–549.

Oostrom, J. K., Born, M. P., Serlie, A. W., & Van der Molen, H. T. (2011). A multimedia situational test with a constructed-response format: Its relationship with personality, cognitive ability, job experience, and academic performance. *Journal of Personnel Psychology, 10*(2), 78–88.

Oostrom, J. K., Born, M. P., Serlie, A. W., & van der Molen, H. T. (2012). Implicit trait policies in multimedia situational judgment tests for leadership skills: Can they predict leadership behavior? *Human Performance, 25*(4), 335–353.

Pang, N., Zhao, X., Wang, W., Xiao, W., & Guo, D. (2021). Few-shot text classification by leveraging bi-directional attention and cross-class knowledge. *Science China Information Sciences, 64*(3).

Qi, S. Q., & Dai, H. Q. (2003). The property function and the development of situational judgment tests. *Psychological Exploration, 23*(4), 42–46.

Ramesh, D., & Sanampudi, S. K. (2022). An automated essay scoring systems: A systematic literature review. *Artificial Intelligence Review, 55*(3), 2495–2527.

Ramineni, C., Trapani, C. S., Williamson, D. M., David, T., & Bridgeman, B. (2012). Evaluation of the e-rater® scoring engine for the GRE® Issue and Argument prompts [EB/OL].

Robson, S. M., Jones, A., & Abraham, J. (2007). Personality, faking, and convergent validity: A warning concerning warning statements. *Human Performance, 21*(1), 89–106.

Rogers, W. T., & Harley, D. (1999). An empirical comparison of three- and four-choice items and tests: Susceptibility to testwiseness and internal consistency reliability. *Educational and Psychological Measurement, 59*(2), 234–247.

Rudner, L. M., & Liang, T. (2002). Automated essay scoring using Bayes' theorem. *The Journal of Technology, Learning and Assessment, 1*(2), 1–22.

Slaughter, J. E., Christian, M. S., Podsakoff, N. P., Sinar, E. F., & Lievens, F. (2014). On the limitations of using situational judgment tests to measure interpersonal skills: The moderating influence of employee anger. *Personnel Psychology, 67*(4), 847–885.

Süzen, N., Gorban, A. N., Levesley, J., & Mirkes, E. M. (2020). Automatic short answer grading and feedback using text mining methods. *Procedia Computer Science, 169*, 726–743.

Tavoosi, S. (2022). *Development and Validation of a Counterproductive Work Behavior Situational Judgment Test With an Open-ended Response Format: A Computerized Scoring Approach* (Unpublished master's thesis). University of Central Florida.

Wang, Y., & Peng, H. L. (2019). Validation on automatic scoring for open-ended questions in Chinese oral tests. *China Examinations, 9*, 63–71.

Weekley, J. A., & Ployhart, R. E. (2005). Situational judgment: Antecedents and relationships with performance. *Human Performance, 18*(1), 81–104.

Whetzel, D. L., & McDaniel, M. A. (2009). Situational judgment tests: An overview of current research. *Human Resource Management Review, 19*(3), 188–202.

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.

Xie, X. Q. (2013). Validation: From reasonable to plausible interpretation of test score. *China Examinations, 7*, 3–8.

Xu, J. P. (2004). *Research on Teacher Competency Model and Evaluation* (Unpublished doctoral dissertation). Beijing Normal University.

Yang, L., Xin, T., Luo, F., Zhang, S., & Tian, X. (2022). Automated evaluation of the quality of ideas in compositions based on concept maps. *Natural Language Engineering, 28*(4), 449–486.

Zhang, Y., Lin, C., & Chi, M. (2020). Going deeper: Automatic short-answer grading by combining student and question models. *User Modeling and User-Adapted Interaction, 30*(1), 51–80.

Zhao, Y., Shen, Y., & Yao, J. (2019, August). Recurrent neural network for text classification with hierarchical multiscale dense connections. *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, 5450–5456, Macao, PEOPLES R CHINA.

Author Contributions Statement:

XU Jing: Conceptualization, methodology, instrument development, data collection/cleaning/analysis, original draft preparation
LUO Fang: Methodology refinement
MA Yanzhen: Model experimentation and comparison
HU Luming: Final manuscript revision
TIAN Xuetao: Model experimentation and comparison, final manuscript revision

Note: Figure translations are in progress. See original paper for figures.