Discover Education

RESEARCH Open Access

A comprehensive item pretesting study to develop an automated short-answer grading platform for Turkish inputs

Merve Kılıç[1,2], Gürkan Cüvitoğlu[1], Şevket Polan[1], Yusuf Balcı[1], Sungur Gürel[2], Elif Kübra Demir[1], Soner Akşehirli[1], Bahadır Namdar[1], Nuri Karakaş[1], Senem Kumova Metin[3], Hakan Atılgan[1], Tarık Kışla[1] and Burak Aydın[1,4*]

*Correspondence:
Burak Aydın
burak.aydin@leuphana.de; burak.aydin@ege.edu.tr
[1]Ege University, İzmir 35100, Türkiye
[2]Siirt University, Siirt, Türkiye
[3]İzmir Economy University, İzmir, Türkiye
[4]Leuphana University, Lüneburg, Germany

Abstract
This study presents several pretesting approaches for automated short-answer grading (ASAG) items, addressing a critical gap in standardized item development procedures in the AS literature. We utilized a large language model approach to develop and assess an ASAG platform for Turkish language inputs across three subject domains: history of the Turkish revolution, Turkish as a first language, and science literacy. Our methodology integrated traditional classical test theory item parameters with insights from cognitive interviews and expert-based machine score-ability ratings to predict ASAG performance. We evaluated performance using quadratic weighted kappa across 73 short-answer items. Our preliminary results indicated that the effectiveness of the pretesting metrics varies by domain. While strong correlations between item parameters and ASAG performance were observed in the history and Turkish domains, a more complex case arose in the science literacy domain, where pretest metrics were insignificant predictors, likely due to linguistic and semantic diversity and complex reasoning. Preliminary evidence suggests that, depending on the study domain, ASAG performance might be predicted from pretesting results. Overall, this study has the potential to advance the field by shifting the paradigm from traditional, retrospective algorithmic validation to a proactive, evidence-based process of item pretesting and development for ASAG.

Keywords Automated short-answer grading (ASAG), Item pretesting, Validity, Machine score-ability, Natural language processing

1 Introduction

Test items can be broadly categorized as selected-response (SR) or constructed-response (CR) [40, 69]. When carefully prepared, CR items have the potential to measure knowledge and skills that cannot be adequately assessed with SR items [16, 31, 46, 72, 75]. There has been an increase in the use of CR items due to the interest in measuring higher-order thinking [78]. Because these items have been utilized extensively across subject domains, the terminology used to define CR items has diversified over time [39,

84]. Historically, terms such as free response, extended response, performance assessment, and even essay have been used to describe this item type. Among the many CR item types, the focus of this study is on short answer (SA) items, in which the test takers are expected to give a short answer, ranging in length from a word, a symbol, a number, an expression, a simple sentence, or a sentence to a paragraph [17]. For instance, a low-cognition SA task might merely require the recall of an explicit historical date or scientific term, whereas a high-cognition SA task can require an examinee to synthesize a brief causal mechanism or evaluate an experimental relationship in a few sentences. Thus, when carefully developed, SA items evaluate complex reasoning without requiring extensive text generation. As with other item types, standardization is important for SA exams; however, these items are generally scored by human raters, which poses a challenge. SA items are generally costly, time-consuming, and challenging to implement, yet they offer important validity advantages. The lack of standardization in SA items, on the other hand, can jeopardize the validity advantage [3, 71], thereby making it challenging to ensure that test results are fair and comparable. This lack of standardized scoring can induce a negative washback effect, where instructors abandon CR items in favor of SR simply to avoid scoring logistics, ultimately narrowing the scope of classroom instruction [1, 70]. To combine the advantages of SA items with the practical need for standardization, this study introduces a comprehensive framework for designing items intended for ASAG. Building on the machine score-ability (MSA) perspective of Lottridge et al. [67], we argue that the effectiveness of an ASAG platform depends not only on the scoring algorithm but also on the linguistic quality and MSA ratings assessed during the item pretesting phase. While traditional guidelines emphasize the importance of training and clear rubrics to ensure reliability among human raters, the rise in the use of ASAG requires a proactive shift in how we pilot these tasks.

This paper specifically evaluates several pretesting approaches across three distinct subject areas: history of the Turkish revolution (HTR, course name: *Atatürk İlkeleri ve İnkılap Tarihi*), Turkish as a first language (TL1), and science literacy (SCL). We use a large language model (LLM)-based approach to score Turkish inputs. Our investigation focuses on how classical test theory parameters—such as difficulty and discrimination—along with findings from cognitive interviews and MSA ratings, serve as predictors of actual ASAG performance, measured using quadratic weighted kappa (QWK). By clarifying these relationships, this study seeks to move beyond typical post hoc algorithmic evaluations. It attempts to establish upfront pretesting steps that secure psychometric and structural item validity prior to operational ASAG deployment.

1.1 A brief review of current SA assessment in practice

Guidelines for preparing SA items (e.g., [8]) emphasize that development should rely on data from samples that reflect the examinee or domain expert population, while ensuring full transparency regarding the purpose, content, and scoring criteria for both community stakeholders and test-takers. SA items should focus on students' cognitive processes, be well-defined tasks, and have the scoring process clarified through rubrics. Gitomer [37] outlined seven principles for effective SA item writing: the use of this item format must be justified, conclusions drawn from student responses should be based on direct information rather than on an implicit structure, scoring levels should be clearly distinguished, each scoring level should have a clear rationale, responses should not be

overly detailed to ensure that interpretations are sufficient for different responses, items should be empirically tested and revised if necessary, and the cognitive structures measured in all tasks should be consistent.

One of the most important justifications for the use of SA items in exams is that they may allow for a more sensitive measurement of students' mastery of the subject [7, 64, 97]. Compared to SR item types, they can reduce guessing effects, which might be difficult to separate from true score variance [36, 98]. However, they may have the disadvantage of reduced content coverage due to fewer items within a test administration [115]. Known limitations of SA include their sensitivity to poor writing skills among test takers [60] and to scoring-related issues such as low reliability despite efforts [89, 115]. These limitations have long been a problematic area in large-scale assessments worldwide [50, 52, 117]. Research shows that various human-rater factors can influence scoring variability and reliability in SA items [9, 52]. These factors include raters' professional backgrounds, scoring methods, scoring criteria, tolerance for grammatical errors, and the number of raters involved in the scoring process [126]. As scoring methods, holistic rubrics emerged primarily to ensure a higher level of inter-rater reliability [15, 53] when compared with analytical rubrics [51]. Nevertheless, both scoring methods have advantages and disadvantages for use in the assessment process, and both require substantial effort to train raters [10, 56]. ASAG systems have the potential to address this challenge [104].

1.2 A brief review of ASAG

By definition, ASAG refers to the use of statistical and computational linguistic techniques to assign scores or labels to responses to open-ended test items [66]. The application of ASAG to complex, structured responses has firmly entered the domain of applied measurement, carrying significant implications for high-stakes decision-making [119]. Prominent examples of this technology include the engines employed by the Educational Testing Service (ETS), such as *e-rater* for essays, *m-rater* for mathematics, *SpeechRater* for speech, and *c-rater* for content. Specifically, the *c-rater* engine is designed to score the content of SA items by evaluating answers against pre-identified correct responses and semantically equivalent expressions. ASAG research for non-English languages has also been gaining momentum (e.g., [18, 24, 55, 59]). Broadly, approaches to ASAG fall into four main categories [21, 41, 43, 54, 65]: feature-based models such as *c-rater*, simple statistical models, machine learning approaches, and deep learning or LLMs.

The development and deployment of an ASAG process typically follow a six-step framework [66]. First, responses representing the target population are collected, after which trained experts assign manual scores using well-defined evaluation criteria [124]. From this manually scored data, at least two subsamples are created to facilitate model comparison; competing models are developed and evaluated, and the best-performing model is selected. Because scoring engines involve many parameters, they are particularly prone to overfitting [107]. Consequently, the selected model must be evaluated on a held-out validation dataset to assess its performance on unseen data before it is packaged and deployed as an operational engine in real-world test environments.

To select and validate the final model, QWK is commonly employed as an evaluation criterion [120], despite its limitations, particularly for small sample sizes [63]. Validation generally relies on comparisons with scores from human raters, who serve as the gold

standard [29, 90, 93, 120]. However, this validation is often addressed indirectly through random sampling, assuming that all score levels are represented equally [26]. A critical limitation is that the scores used to validate AG systems largely depend on the quality of human scoring [92]. Despite potential issues such as errors, costs, inconsistencies, and functional limitations, expert human raters remain the best source for developing AG algorithms [19, 102, 119]. Therefore, even after validation, ASAG is often used as a second score alongside a human rating in high-stakes contexts, ensuring that human input remains essential to the process [29].

Operational evaluation focuses on identifying differences between humans and ASAG to assess the impact of these discrepancies on validity [119]. This process involves four sequential actions: determining human and ASAG evaluations, identifying agreements and discrepancies, selecting cases that are inconsistent to determine the cause, and deciding whether to adopt a human, automated, or compromise perspective. If necessary, the evaluation criteria are modified, and all responses are reprocessed. This rigorous focus on discrepancies clarifies the differences between the two scoring methods and provides crucial information on the accuracy of ASAG. Because human raters will continue to play an important role in the SA scoring process, the procedures underlying this process must be continuously supported by empirical evidence [123].

In general, discussions of ASAG system validation have been addressed indirectly in the academic literature [92]. While machine scoring of essays and SA items is increasingly accepted in the United States, it is noteworthy that these systems are at different stages of development across countries and languages [105]. This situation highlights the need not only to address technical issues but also to provide ongoing evidence demonstrating whether ASAG accurately reflects the qualities of good input [105]. The proliferation of distributed online scoring applications in recent years has transformed large-scale ASAG [116]. However, human raters will remain integral to the process for the foreseeable future [124]. Therefore, careful planning and execution of pilots to ensure the reliability of automated systems and to solidify the role of human raters are critical steps in advancing this field.

Despite criticisms that scoring engines do not genuinely understand language and can be tricked into assigning high scores [86], the rapid development of computer technology has led to increased use of ASAG, which offers distinct advantages over human raters, including the provision of detailed feedback, ensuring consistency and reproducibility, standardizing the application of scoring criteria, and making the scoring rationale traceable. Furthermore, ASAG may enhance objectivity, reliability, and efficiency [118] and enable remote administration and online evaluation [119]. These advantages are particularly vital for SA items, which are traditionally complex and costly to score; thus, computer-based assessments effectively enable the broader use of these item types [57]. A recent systematic review by Cüvitoğlu et al. [21] reported results from 110 ASAG studies and discussed the increased interest in these technologies. However, one of the main findings of this review was that, despite the high research activity, none of the studies reported details of SA development or item pretesting.

1.3 A brief review of SA item pretesting practices

Several stages of the test development process also apply to the ASAG process. Standards for achievement test development have been recommended by widely recognized

organizations, including the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME) [4, 40, 113]. These stages include defining the test's characteristics; item writing and review; item pretesting; reliability and validity analyses; constructing and evaluating the test form; developing administration and scoring procedures; test revision; gathering evidence; and ensuring validity. Şahin et al. [111] examined numerous studies on test development within the framework of these standards. The review included tests with multiple-choice, open-ended, and mixed items. Their findings revealed that most studies did not create an item pool, did not seek input from language experts, did not conduct sufficient pretesting, did not perform item analyses, did not provide most reliability and validity evidence, often relied on expert opinion for validity evidence, did not give information about experts, and did not report the test administration conditions in most cases. However, for example, scoring criteria must be field-tested to ensure that the criteria at each level reflect the intended performance range and to confirm that trained raters yield consistent results [28]. Similarly, Baldwin et al. [8] suggest that all performance tasks and instructions should be piloted with a sample representative of the examinees who will take the operational assessment. Schmeiser and Welch [101] recommend administering approximately twice the number of items ultimately required. Ideally, items should be reviewed by independent experts for linguistic complexity, fairness, accuracy of the answer key, cognitive demand, and alignment with specifications. For new item formats, developers may employ think-aloud protocols to assess accessibility and cognitive engagement. Items should generally be examined for difficulty, discrimination, inter-item relationships, and potential differential item functioning, as described by Wolf et al. [122], who classified SA as cognitively demanding and reported these analyses. Furthermore, rater consistency requires review. If pilot testing is infeasible, Baldwin et al. [8] suggest that raters generate responses addressing seven key areas, specifically determining whether examinees understand directions, whether tasks are appropriate for the group, and whether any bias exists that might advantage specific examinees. Additionally, raters must assess whether tasks elicit the intended response types, whether scoring can be conducted accurately and easily, whether they are using the scoring system in the intended way, and whether rater scoring remains consistent.

During item pretesting, participants can be asked to perform tasks in the researchers' immediate presence to capture real-time behavioral data [25]. This direct observational approach allows researchers to monitor physical indicators of confusion, track behavioral patterns, and document unrecorded modifications such as skipped items or erased and deleted responses [99]. A pilot study with 12 to 50 participants might be sufficient to identify any problems with the test and is cost-effective in terms of time and finances [103, 109]. Hogan [47] recommended conducting an informal pilot study with 10 participants first. This informal study asks participants to think aloud and answer the questions. This prevents the use of ambiguous items in the pilot study and their inclusion in the statistical analysis. A formal pilot study representative of the target population can then be conducted for statistical analysis [47]. The Program for International Student Assessment (PISA), a well-known large-scale assessment, employs a slightly different strategy. Their process begins with an internal expert review of item features, followed by small-scale cognitive interviews using think-aloud protocols to map examinee response

processes, the refined item pool undergoes large-scale field testing with representative samples to stabilize initial psychometric parameters before operational deployment [6]. Recently, PISA conducted an even more rigorous pilot study to assess the invariance of item parameters, estimate the initial parameters of newly added mathematics topics, evaluate survey applications with sampling, and assess the effectiveness of the computer-based application [82]. However, these rigorous item pretesting studies are generally impractical for low-stakes assessments.

In contrast to the rigorous protocols established for traditional assessments, the development and pretesting of items specifically for ASAG appear significantly less standardized. As mentioned above, a recent systematic review by Cüvitoğlu et al. [21] highlighted this gap, reporting that among 110 studies on ASAG, none provided detailed accounts of item development or pretesting procedures. This absence persists despite attempts to formalize such processes; notably, Lottridge et al. [67] proposed an MSA perspective designed to predict automated scoring performance prior to the resource-intensive engine training phase. This framework used expert ratings of specific item features—such as the number of concepts, linguistic variation, and item-rubric alignment—to forecast the likelihood that an engine would yield scores comparable to those of human raters. However, the predictive validity of this scale was dependent on the item domain. Although citations of this work acknowledge the importance of these item features (e.g., [125]), the specific rating scale has not been widely adopted as a standard quality control mechanism. On the other hand, recent studies have investigated the potential of LLMs for item pretesting (e.g., [13, 30, 45, 48, 58]). It is stated that by feeding item stems and options into an LLM, item parameters can be estimated based on characteristics such as textual complexity, readability, stem-option similarity, and semantic structure; these estimates can then be used as informative priors, particularly in small-sample calibration studies, enabling a reduction in the required sample size [2, 11, 49, 68, 94, 114]. However, the literature emphasizes that LLMs should be used as a complementary support system under human supervision, rather than as a replacement for human expertise [79, 85, 106]. Nevertheless, the field largely relies on retrospective empirical analysis rather than proactive item pretesting, leaving the MSA of items as an unverified assumption until after data collection.

1.4 Current study

The current study was conducted as part of a project funded by the Scientific and Technological Research Council of Türkiye (TÜBİTAK). The project aimed to develop an ASAG approach for Turkish inputs in the subject domains of HTR, TL1, and SCL, and to identify factors affecting agreement between ASAG and experts. Acknowledging the paucity of information on item pretesting for ASAG in the literature, we have comprehensively reviewed approaches to item pretesting. Specifically, for each of the three subject domains, we conducted traditional statistical item pretesting analyses using a representative but relatively small sample, conducted cognitive interviews, and asked four experts about the MSA of the developed items [67]. After this comprehensive item pretesting step, developing the ASAG platform and collecting data to evaluate its performance took approximately 17 months. ASAG's performance or observed score-ability for the study SA items in the three different subject domains was evaluated using QWK.

To complement our analysis, we also fed our items and rubrics into an LLM and asked it to assign an MSA score and predict the QWK.

The presupposition of linking pretest parameters to automated scoring performance resides in response variance and rubric clarity [67, 120]. The item difficulty might be related to the linguistic and semantic diversity of the questions [44]. Extremely easy or difficult items probably restrict score variance [40] and compress response variations into possibly predictable lexical clusters (e.g., uniform correct keywords or omitted answers), easing machine alignment. Moderate-difficulty items, however, might maximize linguistic variations and partial misconceptions, challenging an LLM's boundary-discrimination capabilities [21]. Similarly, item discrimination reflects how successfully an item or its rubric separates distinct levels of the latent construct [101]. Highly discriminating items might possess clear, unambiguous semantic boundaries between score tiers [30, 90], maximizing human-machine alignment (e.g., QWK). Conversely, poorly discriminating items often suffer from construct irrelevance [92] or rubric ambiguity, confounding human experts and LLMs alike. Therefore, variation in human cognitive processing and response production might systematically modulate both the item parameters and the subsequent algorithmic score-ability of the text inputs. The paths representing the presupposition of relationships among item attributes, cognitive processing, psychometric parameters, and AS performance are conceptualized in Fig. 1. Following this presupposition, the current study's research questions were formed separately for the three subject domains as follows:

RQ1: How well do traditional item parameters from the item pretesting correlate with the observed QWK across items?

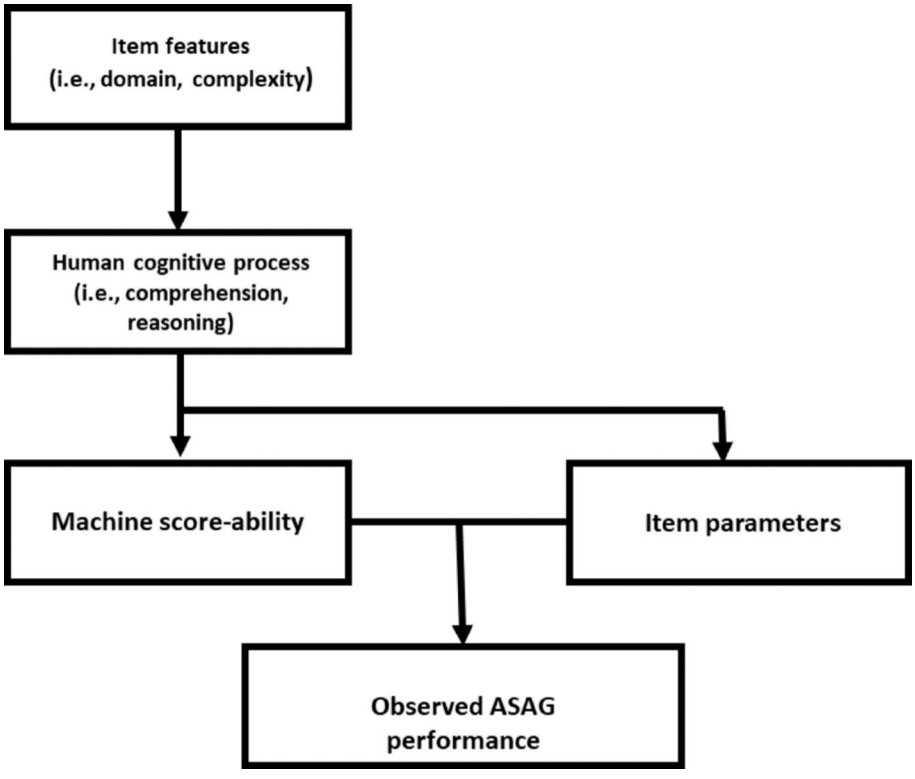

Fig. 1 The presumed mechanisms linking item parameters to AS performance

RQ2: How well do MSA ratings and predicted QWK correlate with the observed QWK across items?

These research questions were inspired by and similar to those stated in Lottridge et al. [67]. For the SCL domain only, we formed an additional RQ as follows:

RQ3: How well do perceived item difficulty and response process quality from the cognitive interviews correlate with the observed QWK across SCL items?

2 Method

In this section, we introduce: (a) the subject domains and experts, (b) items, rubrics, and pretesting approaches, (c) examinee sample information, (d) ASAG for Turkish and its evaluation measures, and (e) outcome evaluation measures.

2.1 Subject domains

The HTR course was first introduced in 1925 as History of Revolutions in order to analyze the Turkish Revolution comparatively. It became a regular university offering in 1933 with the establishment of the Institute of History of the Turkish Revolution at Istanbul University. As of May 27, 1960, it became a mandatory two-semester course. Under Higher Education Law No. 2547 (November 6, 1981), it was adopted as a compulsory common course for all undergraduate programs. The course aims to enable students to evaluate Ottoman-Turkish modernization, grasp the military and diplomatic phases of the War of Independence, and interpret core concepts such as National Sovereignty. Learning outcomes include analyzing the foundation of the Republic, the transition to a democracy, economic development challenges, and Türkiye's strategic global position. An experienced faculty member who serves as the department head and the college coordinator of the HTR course served as the subject expert.

Our TL1 definition distinguishes between native language education—specifically, Turkish—and foreign language learning. While native language is a core cognitive ability acquired naturally, generally at an early age, foreign language learning involves the conscious acquisition of new structures. Turkish instruction generally covers five skill areas, centralized exams generally focus only on reading and grammar. Modern reading instruction prioritizes analytical and critical reading. The PISA 2018 reading framework categorizes these into three cognitive processes: accessing & retrieving information, integrating & interpreting, and reflecting & evaluating. Accordingly, PISA defines reading literacy as understanding, using, and reflecting on written texts to achieve one's goals and participate in society [83]. In this context, SA questions are deemed highly suitable for measuring these skills. An experienced faculty member who is also a board member of the Turkish Language Association, the regulatory body for the Turkish language, served as the TL1 expert.

SCL has long been a core objective in national and international education policies [5, 73, 74, 77]. PISA defines SCL as the competence to use scientific knowledge to address daily problems and make informed decisions [80], a goal mirrored by the Turkish Higher Education Qualifications Framework's emphasis on evidence-based problem solving. Climate change serves as a critical interdisciplinary context for developing these skills across various undergraduate courses (e.g., [12, 62, 110]). Consequently, we used the PISA framework, which categorizes science literacy into three dimensions: scientific knowledge (content, procedural, and epistemic), competencies (explaining phenomena,

evaluating research, and interpreting data), and attitudes (interest, valuing science, and environmental awareness) [81]. Research on climate education typically focuses on topics such as greenhouse gases, carbon footprints, and human impacts on climate (e.g., [23, 62, 76]). For the current study, a faculty member with substantial experience and an academic publishing record on the topic served as the SCL expert. In addition to subject-domain experts, the ASAG for the Turkish platform project included four computer scientists: two faculty members with substantial publication records in natural language processing and two graduate students with extensive programming experience. These four project members served as the expert panel to answer the MSA scale.

2.2 Items, rubrics, and pretesting approaches

The study followed the same item and rubric preparation across the three domains: HTR, TL1 and SCL. For each domain, an initial pool of 50 SA items, along with their scoring rubrics, was prepared. Scoring rubrics included instructions to assign a score of 0 for completely incorrect, 1 for partially correct, or 2 for correct answers. For all three domains, student responses were collected from a representative sample to estimate item parameters using classical test theory, with human expert scoring. Specifically, we computed the average item score, which ranged from 0 to 2, with higher scores indicating items of lower difficulty. We also computed the Pearson correlation coefficient (r) between each item and the total score to quantify item discrimination, with higher correlation values indicating better discrimination. Based on these parameters, 25 SA items per domain were selected for further pretesting.

In the second step of item pretesting, we shared items and rubrics with the computer experts and independently asked them to rate score-ability based on the five descriptors proposed by Lottridge et al. [67]. Specifically, a rating of 5 indicates that the ASAG can yield scores comparable to those of domain experts, whereas a rating of 1 indicates that items are not suitable for the ASAG. We also requested a brief explanation for any rating of 3 or lower. To complement this pretesting approach, we asked for Gemini's [34] MSA ratings and its QWK prediction.

Only for the SCL domain, the expert conducted a cognitive interview study. A total of 3 students (2 male, 1 female) in the undergraduate science education program and 1 student (1 female) in the graduate program participated in individual interviews. All interviews were audio-recorded with the participant's consent and transcribed verbatim by the expert. Each interview lasted approximately 30 min. First, the expert explained the purpose of the interview and provided the participants with the 25 questions in writing. For each question, participants were asked to answer it, state their level of confidence, and identify its level of difficulty. Additionally, at the end of the interview, each participant was asked whether the 25-question set reflects scientific literacy in the context of climate change and whether it is suitable for identifying high- and low-achieving students' scientific literacy. For each answer, we conducted content analysis following Collins [20] using a combined think-aloud and verbal probing approach to assess participants' understanding of the questions and response generation processes. We applied process-quality analysis to examine participants' question processing across 4 aspects: (a) their comprehension of the key terms and the question as a whole, (b) their retrieval of information by identifying whether respondents can recall the required information, (c) whether the respondent used strategies to make judgments about the question

and self-reported confidence as an indicator of certainty, (d) whether, in their response, the reasoning chain was internally consistent and aligned with the question's intent. Perceived item difficulty was treated as a separate descriptive variable, noted as easy, medium, or difficult, and summarized by coding these categories as 1, 0.5, and 0, respectively, and averaging across participants. We placed greater analytical emphasis on comprehension to reduce ambiguity and ensure interpretability, which was critical for both item pretesting and automated scoring. Accordingly, comprehension was operationalized in two subcomponents (key terms and the item as a whole), yielding five scored components in total (the two comprehension subcomponents plus retrieval, judgement, and response). Each component was scored using a 0–2 rubric (0, does not meet criteria; 1, partially meets criteria; 2, fully or nearly fully meets criteria), and component scores were summed to yield a response processing score (RPS) ranging from 0 to 10. Overall, the item set showed an acceptable RPS (6.48/10), and perceived difficulty ratings were distributed across easy (29%), medium (38%), and difficult (33%). Questions with the lowest process-quality totals were also perceived as the most difficult, supporting convergence between observed processing and perceived difficulty. In particular, three items were consistently rated as difficult and often left unanswered, which participants attributed to unfamiliar or advanced concepts. Finally, participants' evaluative comments supported the content validity of the set for scientific literacy in the climate change context and its discriminant power for distinguishing higher- and lower-achieving students.

As the final step prior to data collection to evaluate ASAG, the domain experts studied the content validity of the selected 25 items. The HTR and SCL experts each decided to replace one item, whereas the TL1 expert approved all the selected 25 items.

2.3 Examinee sample information

For the SCL domain, the representative sample and the main study sample comprised 109 and 387 college students, respectively, yielding 15,125 scored responses. As explained earlier, four college students participated in the cognitive interview, which was conducted between the two data collection periods. For TL1, the representative sample comprised 94 students, whereas the main study included 491 students, yielding 16,975 student responses. For HTR, data from 488 students formed the representative sample, whereas the main study included 780 students and yielded 43,900 responses. Overall, the majority of participants were from Ege University across diverse majors between 2023 and 2025; a limited number of responses were gathered from three additional universities within Türkiye to increase the amount of data collected.

2.4 ASAG for Turkish and its evaluation

A zero-shot evaluation approach was used to assess student responses, leveraging LLM capabilities. This method involved applying evaluation criteria via prompt engineering, enabling the model to score responses as a human expert would [61]. The system consisted of four stages: data preparation, dynamic prompt generation, querying via the Gemini API, and reporting structured outputs. The dataset included question texts, expert-prepared reference answers, and student responses, all of which underwent cleaning and preprocessing. This served as the input architecture for the LLM, with no structural changes. The prompt engineering process was crucial for defining the task. A dynamic prompt configuration was created for each response, encompassing role

assignment, context integration, and scoring logic [27]. The model was assigned a specific role with domain knowledge to ensure that evaluations met academic standards. Context integration brought together the question, student answer, and expert rubric. The scoring logic involved either searching for reference-answer matches or using the model's expertise when none were available. The Gemini 1.5 Flash model was chosen for its strong contextual understanding, high reasoning capacity, and low latency [35, 112]. To ensure reliable scoring, the temperature was set to 0.2 and the *topP* to 1, minimizing randomness. To facilitate analysis, the Gemini API's JSON Mode was enabled, producing outputs in a fixed JSON schema that included scores and rationales. The entire process was automated using a Python-based system that managed responses, communicated with the Gemini API, and integrated data into a central database. An advanced error handling and retry mechanism was implemented to mitigate potential data loss during high-volume processing, thereby ensuring system stability and data integrity. The performance of our system was evaluated by computing QWK between ASAG and human expert scores; an item was flagged as successfully scored by ASAG when QWK > 0.65 [87].

2.5 Outcome evaluation measures

To address our research questions, following Lottridge et al. [67], we computed correlations between the ASAG evaluation metric and pretest measures for 25 TL1 items and 24 HTR and SCL items [42, 88, 91, 96]. Specifically for all three domains, pretest measures were (a) item difficulty and discrimination values based on the data from a representative sample and human expert scores using classical test theory, (b) MSA ratings, and (c), only for the SCL domain, the average perceived difficulty based on the cognitive interviews. When interpreting the magnitude of these correlations, readers should note that with a small item-level sample size, coefficients possess wider confidence intervals and represent preliminary effect sizes that require replication in larger item pools rather than highly robust population parameters.

3 Results

For the HTR domain items, difficulties ranged between 0.05 and 1.77, with a mean of 1.15 out of 2, and the correlation between the item difficulty and QWK was positive and statistically significant ($r_{\text{diff–QWK}} = 0.67$, $p < 0.001$). The discrimination values ranged from 0.06 to 0.46, with a mean of 0.28, and $r_{\text{disc–QWK}} = 0.68$, $p < 0.001$. Expert MSA ratings were generally high across all items, resulting in average MSA values between 4.25 and 5, with a mean of 4.73 and $r_{\text{MSave–QWK}} = 0.69$, $p < 0.001$. Gemini's MSA ratings had an average of 4.46, slightly lower than the experts' average, but its correlation with the observed QWK was noticeably larger with $r_{\text{MSgem–QWK}} = 0.833$, $p < 0.001$.

The ASAG for HTR was considered successful, with an average QWK of 0.76; 18 of 24 items (75%) had a QWK greater than 0.65. Gemini's predicted QWK based on the items and rubrics ranged from 0.55 to 0.95 with a mean of 0.86; the correlation between the predicted and observed QWK was strong, $r_{\text{QWKgem–QWK}} = 0.85$, $p < 0.001$. The experts' MSA average was 4.82 for items with QWK > 0.65 and 4.46 for the remainder; this difference was statistically significant, Welch's $t(6.34) = 2.51$, $p = 0.04$, with a Cohen's d of 1.21. Gemini's MSA average was 4.83 for items with QWK > 0.65 and 3.33 for the remainder,

Welch's $t(5.47) = 3.48$, $p = 0.02$, with a Cohen's d of 1.93. Pretest and ASAG measures for the HTR domain are reported in Table 1.

For the TL1 domain items, difficulties ranged between 0.64 and 1.80, with a mean of 1.04 out of 2, and $r_{\text{diff–QWK}} = 0.55$, $p < 0.01$. The discrimination values ranged between 0.04 and 0.45, with a mean of 0.25, and $r_{\text{disc–QWK}} = 0.07$. MSA ratings were generally high across all items, yielding average MSA values between 3.75 and 5, with a mean of 4.47 and $r_{\text{MSave–QWK}} = 0.27$. Gemini's MSA ratings averaged 3.96, with $r_{\text{MSgem–QWK}} = 0.00$.

The ASAG for TL1 was also considered successful, with an average QWK of 0.70; 17 of 25 items (68%) had a QWK greater than 0.65. Gemini's predicted QWK had a mean of 0.83 with $r_{\text{QWKgem–QWK}} = 0.08$. The experts' MSA average was 4.54 for items with QWK > 0.65 and 4.31 for the remainder; this difference was not statistically significant, Welch's $t(15.7) = 1.61$, $p = 0.13$ with a Cohen's d of 0.67. Gemini's MSA ratings were similar across two groups, Welch's $t(10.9) = 0.61$. Pretest and ASAG measures for the TL1 domain are reported in Table 2.

For the SCL domain items, difficulties ranged between 0.34 and 1.39, with a mean of 0.73 out of 2, and $r_{\text{diff–QWK}} = -0.06$. The discrimination values ranged between 0.26 and 0.69, with a mean of 0.42, and $r_{\text{disc–QWK}} = 0.16$. Average expert MSA ratings were between 2.25 and 4.50, with a mean of 3.30 and $r_{\text{MSave–QWK}} = 0.20$. Gemini's MSA ratings averaged 3.33, with $r_{\text{MSgem–QWK}} = 0.35$. Perceived difficulty and RPS from the cognitive interviews had a mean of 0.48 and 6.30, respectively, and were strongly correlated ($r_{\text{Pdiff,–RPS}} = 0.92$, $p < 0.001$). The correlation between observed difficulty and perceived item difficulty was $r_{\text{diff, Pdiff}} = 0.66$, $p < 0.001$; similar to the correlation between observed

Table 1 Pretest and ASAG measures for the HTR domain

Item ID	Diff.	Disc.	MS_{R1}	MS_{R2}	MS_{R3}	MS_{R4}	MS_{Ave}	MS_{Gem}	QWK_{Gem}	QWK	QWK > 0.65
HTR 1	1.54	0.26	5	5	5	5	5.00	5	0.94	0.99	1
HTR 3	1.34	0.24	5	4	5	5	4.75	5	0.90	0.78	1
HTR 4	1.55	0.28	4	5	5	5	4.75	5	0.95	0.98	1
HTR 5	1.49	0.30	5	5	5	5	5.00	5	0.95	0.99	1
HTR 6	1.77	0.38	5	5	5	5	5.00	5	0.95	0.97	1
HTR 7	1.62	0.22	4	5	5	4	4.50	4	0.78	0.72	1
HTR 8	1.24	0.17	4	4	5	4	4.25	3	0.60	0.50	0
HTR 9	1.05	0.27	5	5	5	5	5.00	5	0.94	0.85	1
HTR 10	1.21	0.39	4	5	5	5	4.75	5	0.90	0.89	1
HTR 11	0.20	0.11	4	4	5	4	4.25	2	0.55	0.38	0
HTR 12	0.05	0.06	4	4	5	4	4.25	3	0.60	0.41	0
HTR 13	1.00	0.46	4	5	5	4	4.50	5	0.94	0.96	1
HTR 14	1.15	0.28	5	4	5	4	4.50	4	0.80	0.66	1
HTR 15	0.65	0.13	5	5	5	5	5.00	5	0.95	0.61	0
HTR 16	1.13	0.26	5	4	5	5	4.75	4	0.80	0.47	0
HTR 17	1.73	0.34	5	5	5	5	5.00	5	0.92	0.90	1
HTR 18	0.84	0.12	5	5	5	4	4.75	5	0.90	0.68	1
HTR 19	1.05	0.29	5	4	5	4	4.50	4	0.82	0.78	1
HTR 20	1.70	0.40	4	5	5	5	4.75	5	0.95	0.80	1
HTR 21	1.13	0.42	5	5	5	5	5.00	5	0.95	0.91	1
HTR 22	0.62	0.33	5	5	5	5	5.00	5	0.95	0.85	1
HTR 23	1.37	0.39	5	5	5	5	5.00	5	0.95	0.73	1
HTR 24	1.49	0.37	5	5	5	5	5.00	5	0.95	0.87	1
HTR 25	0.73	0.30	4	4	5	4	4.25	3	0.65	0.48	0
Average	1.15	0.28	4.63	4.67	5.00	4.63	4.73	4.46	0.86	0.76	0.75

Diff Difficulty, *Disc* Discrimination, *MSA* Machine score-ability, *R* rater, *Gem* Gemini

Table 2 Pretest and ASAG measures for the TL1 domain

Item ID	Diff.	Disc.	MS_{R1}	MS_{R2}	MS_{R3}	MS_{R4}	MS_{Ave}	MS_{Gem}	QWK_{Gem}	QWK	QWK > 0.65
TL1 1	1.00	0.35	4	5	5	5	4.75	5	0.96	0.91	1
TL1 2	1.34	0.41	3	5	5	5	4.50	5	0.94	0.95	1
TL1 3	1.28	0.30	5	5	5	3	4.50	5	0.92	0.69	1
TL1 4	1.39	0.45	4	5	5	5	4.75	5	0.98	0.93	1
TL1 5	0.64	0.19	5	4	5	5	4.75	5	0.95	0.26	0
TL1 6	0.79	0.30	3	4	5	4	4.00	5	0.94	0.48	0
TL1 7	1.80	0.04	5	5	5	5	5.00	5	0.96	0.94	1
TL1 8	1.24	0.21	5	5	5	5	5.00	4	0.88	0.74	1
TL1 9	0.97	0.17	3	5	4	5	4.25	4	0.85	0.39	0
TL1 10	1.34	0.40	5	5	5	4	4.75	5	0.95	0.83	1
TL1 11	0.88	0.20	5	4	5	4	4.50	5	0.92	0.35	0
TL1 12	1.17	0.13	4	5	5	5	4.75	4	0.87	0.74	1
TL1 13	0.89	0.18	5	4	5	1	3.75	5	0.97	0.65	1
TL1 14	0.73	0.20	4	4	5	4	4.25	4	0.86	0.82	1
TL1 15	0.79	0.36	4	4	5	2	3.75	4	0.84	0.46	0
TL1 16	1.22	0.42	3	4	5	4	4.00	2	0.65	0.78	1
TL1 17	0.99	0.20	4	4	5	5	4.50	3	0.72	0.80	1
TL1 18	1.05	0.21	4	4	5	3	4.00	3	0.68	0.88	1
TL1 19	0.84	0.28	4	4	4	5	4.25	2	0.58	0.60	0
TL1 20	0.65	0.24	5	4	5	5	4.75	3	0.74	0.91	1
TL1 21	0.96	0.33	4	4	5	5	4.50	2	0.55	0.50	0
TL1 22	0.81	0.23	4	4	5	5	4.50	3	0.71	0.45	0
TL1 23	0.95	0.11	4	4	5	5	4.50	3	0.75	0.91	1
TL1 24	1.21	0.18	5	5	4	4	4.50	4	0.86	0.95	1
TL1 25	1.05	0.27	5	5	5	5	5.00	4	0.82	0.68	1
Average	1.04	0.25	4.24	4.44	4.88	4.32	4.47	3.96	0.83	0.70	0.68

Diff Difficulty, *Disc* Discrimination, *MSA* Machine scoreability, *R* rater, *Gem* Gemini

difficulty and RPS ($r_{\text{diff–CogInt}} = 0.58$, $p < 0.01$). However, the correlation between perceived difficulty and QWK was negative but statistically insignificant ($r_{\text{QWK–Pdiff}} = -0.37$, $p = 0.08$), similar to RPS and QWK ($r_{\text{QWK–RPS}} = -0.34$, $p = 0.11$).

The ASAG for SCL was also considered successful, with an average QWK of 0.70; 18 of 24 items (75%) had a QWK greater than 0.65. Gemini's predicted QWK had a mean of 0.64 with $r_{\text{QWKgem–QWK}} = 0.16$. The experts' MSA average was 3.36 for items with QWK > 0.65 and 3.13 for the remainder; this difference was statistically insignificant, Welch's $t(11.9) = 1.10$, $p = 0.292$, with a Cohen's d of 0.48. Gemini's MSA ratings were similar across two groups, Welch's $t(16.4) = 0.50$. Pretest and ASAG measures for the SCL domain are reported in Table 3.

4 Discussion

Unlike conventional ASAG research that evaluates machine performance retrospectively, this study introduces proactive quality control steps. By validating item parameters and machine score-ability upfront, we sought to bridge the gap between educational measurement and computer science literature, ensuring items are intentionally prepared for AS. Our preliminary findings provide critical insights into the relationships among traditional psychometric item parameters, MSA ratings, and the observed performance of an LLM-based ASAG (QWK ≥ 0.70).

The results across the three subject domains—HTR, TL1, and SCL—indicate that the effectiveness of pretesting for a successful ASAG implementation varies substantially

Table 3 Pretest and ASAG measures for the SCL domain

Item ID	Diff.	Disc.	P_{diff}	RPS	MS_{R1}	MS_{R2}	MS_{R3}	MS_{R4}	MS_{Ave}	MS_{Gem}	QWK_{Gem}	QWK	QWK > 0.65
SCL 1	1.39	0.42	1.00	10.00	3	3	3	2	2.75	3	0.65	0.58	0
SCL 2	1.31	0.45	0.88	9.25	3	4	5	2	3.50	3	0.68	0.64	0
SCL 3	1.18	0.36	0.75	8.25	3	2	5	2	3.00	1	0.10	0.67	1
SCL 4	1.04	0.26	0.88	9.00	4	4	5	3	4.00	3	0.66	0.65	1
SCL 5	0.36	0.41	0.63	8.00	3	3	3	2	2.75	4	0.74	0.61	0
SCL 6	0.97	0.49	0.63	6.50	4	5	5	4	4.50	5	0.84	0.72	1
SCL 7	0.85	0.29	0.88	8.75	3	3	3	3	3.00	4	0.75	0.63	0
SCL 8	0.85	0.35	0.50	6.25	4	4	5	3	4.00	5	0.88	0.79	1
SCL 9	0.88	0.30	0.63	8.75	4	3	4	2	3.25	3	0.68	0.77	1
SCL 10	0.77	0.37	0.50	9.50	5	4	4	2	3.75	3	0.62	0.62	0
SCL 11	0.75	0.43	0.50	6.75	3	4	4	2	3.25	4	0.72	0.77	1
SCL 12	0.83	0.28	0.13	5.00	3	3	3	2	2.75	2	0.48	0.65	1
SCL 13	0.67	0.53	0.63	8.75	3	2	2	3	2.50	3	0.64	0.74	1
SCL 14	0.73	0.68	0.38	5.50	3	3	4	3	3.25	3	0.60	0.79	1
SCL 15	0.67	0.38	0.00	2.50	4	5	5	2	4.00	5	0.86	0.91	1
SCL 16	0.56	0.38	0.63	6.75	3	3	4	3	3.25	4	0.76	0.77	1
SCL 17	0.51	0.46	0.38	6.75	3	3	4	4	3.50	4	0.70	0.70	1
SCL 18	0.58	0.31	0.13	2.50	3	3	1	2	2.25	1	0.15	0.84	1
SCL 19	0.34	0.39	0.33	4.00	4	4	2	2	3.00	2	0.55	0.23	0
SCL 20	0.58	0.69	0.17	1.67	3	3	5	1	3.00	4	0.72	0.77	1
SCL 21	0.48	0.47	0.75	10.00	3	3	4	3	3.25	2	0.45	0.66	1
SCL 23	0.38	0.44	0.00	1.25	3	4	5	2	3.50	5	0.82	0.85	1
SCL 24	0.45	0.42	0.13	3.75	4	4	5	2	3.75	2	0.42	0.78	1
SCL 25	0.37	0.45	0.00	1.75	4	4	4	2	3.50	5	0.85	0.74	1
Average	0.73	0.42	0.47	6.30	3.42	3.46	3.92	2.42	3.30	3.33	0.64	0.70	0.75

Diff Difficulty, *Disc* Discrimination, *MSA* Machine score-ability, *R* rater, *Gem* Gemini, *Pdiff* perceived difficulty

across domains. In the HTR domain, we found strong, statistically significant correlations between the observed QWK, traditional item difficulty ($r = 0.67$), discrimination ($r = 0.69$), expert MSA ratings ($r = 0.69$), Gemini's MSA ratings ($r = 0.83$), and QWK prediction ($r = 0.85$). Compared to TL1 and SCL, HTR items were content-heavy and fact-based. These results show that for such social science domains (e.g., language and science literacy), ASAG developers might want to enhance their rubrics, especially for items with low item difficulty parameters (i.e., low student success), discrimination parameters, and MSA ratings. This supports the idea that clearly defined concepts and distinct boundaries between the correct and incorrect enhance agreement between humans and machines.

The SCL domain presented a more complex scenario compared to the others. Neither the traditional item parameters nor the MSA ratings showed significant correlations with the observed QWK. Through the cognitive interview process, we found a moderate correlation ($r = 0.58$) between perceived difficulty and actual item difficulty; however, observed ASAG performance was not related to either. This discrepancy may be attributed to the nature of scientific reasoning items. As Lottridge et al. [67] noted, while science items may benefit from specific vocabulary, they can suffer if the assessment engine fails to capture the underlying logic or relationships between phenomena. Our preliminary findings suggest that, despite using the same number of question-answer pairs, the higher number of unique words in the alternative answers of SCL domain pairs (SCL: 368, HTR: 263, TL1: 88) indicates greater linguistic and conceptual diversity in the science domain. This type of diversity is challenging for ASAG systems (e.g., [44, 100]) in two ways. Firstly, as the range of expression forms encountered by the model expands, the training data scope becomes relatively sparser, forcing the system to generalize across many surface realizations of similar meanings. Secondly, scientific answers often require multi-level reasoning, precise units and numerical values, and conceptual justification; therefore, even small linguistic differences (e.g., the choice of a technical term, numerical values, units, or causal phrasing) can materially affect scientific accuracy. We also noticed that items that required students to provide a list (e.g., give three examples of the effects of climate change on human health) as the correct response were associated with low ASAG performance. Furthermore, we noticed low ASAG performance when the scoring rubric may not have adequately operationalized the distinction between descriptive statements and mechanism-based reasoning, particularly on cognitively demanding items. Overall, as Bolgova et al. [14] and Grevisse [38] indirectly noted, our results indicate that item- and domain-specific pretesting is necessary.

The implementation of the MSA rating scale proposed by Lottridge et al. [67] produced mixed results. For the HTR and TL1 categories, items successfully scored by ASAG (with a QWK greater than 0.65) received significantly higher MSA ratings than those not scored. This indicates that expert judgment can, to some extent, identify items that possess favorable linguistic or structural features—such as clear rubrics and limited variation in concepts—that lend themselves to ASAG. Despite the low correlation between MSA ratings and observed QWK for the SCL domain, the average MSA ratings for the SCL were noticeably lower than those for the HTR and TL1, indicating that both human experts and an LLM noticed that scoring SCL would be more challenging. Nevertheless, the lack of a significant correlation in the SCL domain aligns with the original MSA study's findings, which reported that expert judgment was moderately

successful in science. Our results suggest that experts may underestimate the capabilities of modern LLMs in managing linguistic variation. In domains like SCL, where responses tend to be more diverse, the "bag-of-words limitations" associated with older scoring engines—which influenced the original 2018 MSA framework—may no longer be as relevant. Nevertheless, our results clearly show the need for new pretesting procedures to predict ASAG performance for our SCL tasks. Our results suggest that rubrics for LLM-based ASAG in science must explicitly define the causal phrasing required for a full-credit score. Future pretesting protocols for complex domains like science should move beyond traditional metrics in order to map the semantic network of acceptable answers and rigorously evaluate how well the rubric operationalizes reasoning prior to administration.

4.1 Limitations and future research

Despite the comprehensive nature of this pretesting study, several limitations must be acknowledged. First, the sample sizes for the representative pretesting phase, which ranged from 94 to 488 students, varied across domains. While these sizes are sufficient for initial estimates based on classical test theory, larger samples would enable the use of item response theory, potentially providing more stable difficulty estimates. Second, the expert panel responsible for the MSA ratings consisted primarily of computer scientists. Although they possess in-depth knowledge of NLP capabilities, incorporating domain-specific experts into the MSA rating process could yield different results, particularly regarding concept overlap and relationships, as described by Lottridge et al. [67]. Third, this study utilized a zero-shot approach with a specific LLM. The performance of automated scoring can be highly sensitive to prompt engineering and the specific model used. Fourth, while we introduce a baseline presupposition linking CTT parameters to AS via cognitive and semantic variance, our cross-domain findings remain partially exploratory. The unique linguistic properties of a specific domain (e.g., science literacy) require further theoretical modeling to successfully predict AS performance. Fifth, while this study evaluates the relational patterns between pretest indicators and AS performance using simple agreement metrics, it does not execute predictive modeling such as multiple regression. Sixth, while the expert MSA ratings were summarized via mean scores, formal inter-rater reliability metrics are not reported. However, to maintain absolute transparency and enable future researchers to assess rater consensus or conduct predictive modeling, the raw rater metrics, item-level parameters, and full datasets have been made accessible. Seventh, while the content-alignment and baseline scoring benchmarks relied on individual subject-matter experts within each domain, these individuals were senior faculty members representing the highest tiers of their respective disciplines within Turkey. Each expert possessed over 15 years of active research experience, held a doctoral degree in the target domain, and maintained documented leadership roles on national curriculum design or ministerial assessment committees. Although relying on a single senior expert per domain is an established psychometric practice for securing baseline content-valid keys, it represents a localized constraint; deploying expanded, multi-member panels of independent reviewers in future iterations could further control for idiosyncratic grading stringency and refine baseline parameter estimation. Eighth, the strong alignment observed between the LLM's QWK predictions and actual agreement performance (e.g., within the HTR domain) must be interpreted with strict

caution. These correlations are exploratory, are not yet operationally validated, and face significant threats to generalizability due to prompt sensitivity and structural dependencies within the scoring rubrics. Furthermore, implementing LLMs as proactive pretesting utilities introduces distinct operational vulnerabilities, including API latency shifts, model drift across version updates, and systemic replication challenges inherent to nondeterministic architectures. Consequently, these preliminary predictions function only as supplemental metadata.

Future research should investigate whether the correlations identified in this pretesting hold true across different architectures. We notice that our complementary analysis, using the same LLM approach we used in the development phase to predict QWK and rate MSA, could be improved by leveraging recent developments and different LLM approaches; however, our results might be regarded as an invitation to conduct more research on LLMs' pretesting capabilities, and readers should notice that this suggestion is consistent with our preliminary findings that LLM's MSA ratings were somewhat better than those of the experts for the SCL domain. Another limitation is that due to time and cost constraints, we conducted cognitive interviews only for the SCL domain, not for HTR and TL1. Finally, while this study provides suggestions for pretesting AS items, it focuses on a single-part response format. As assessments evolve toward more complex, multi-part constructed-response formats, the interplay among item design, rubric complexity, and machine performance will require advanced pretesting strategies. Furthermore, future research can systematically account for individual differences in examinee literacy backgrounds, and baseline reading comprehension skills as potential confounding variables in ASAG performance. In domains like TL1, tasks evaluating explicit morphosyntactic and metalinguistic knowledge assume a healthy, natural baseline variation in literacy skills across the student sample [22, 108]. Within the Turkish linguistic context, recent evidence confirms that individual differences are critical predictors of advanced reading and grammar outcomes [32, 33]. Because our operational grading framework presupposes that observed text variance is purely reflective of target construct mastery rather than underlying reading literacy gaps, omitting these predictor variables remains an administrative limitation.

Appendix

HTR Items

Item ID	SA item
HTR 1	Sultan Abdülmecit döneminde 1856 yılında ilan edilen hangi ferman ile gayrimüslimlere yeni haklar ve bazı ayrıcalıklar tanınmıştır?
HTR 2	Mustafa Kemal Paşa, 28 Mayıs 1919'da yayınladığı genelgede (Havza Genelgesi) yaşanan işgaller karşısında büyük ve heyecanlı mitinglerin yapılarak milli gösteride bulunmasını istemiştir. Mustafa Kemal Paşa, bu genelge ile hangi şehirlerin işgalini örnek göstererek yapılan işgallerin protesto edilmesini istemiştir?
HTR 4	İtilaf Devletlerinin Mondros Mütarekesi'ne dayanarak Anadolu'da başlattıkları işgaller karşısında ortaya çıkan direniş hareketini İttihatçılığın devamı ve İstanbul'daki "meşru" hükümete karşı girişilmiş bir ayaklanma ve başkaldırı olarak gören ve Milli Mücadele karşıtı bir siyaset takip eden Osmanlı sadrazamı kimdir?
HTR 5	İtilaf Devletleri I. İnönü Muharebesi'nin kazanılmasından sonra TBMM hükümetini hangi konferansa davet etmişlerdir?

Item ID	SA item
HTR 6	Atatürk döneminde 1920'li yıllarda uygulanan ekonomi politikaları hangi kongrede alınan kararlarla saptanmıştır?
HTR 7	Kuvayi Milliye hangi anlama gelmektedir?
HTR 8	İzmir Yunanistan yerine başka bir İtilaf devleti tarafından işgal edilseydi, bunun Milli Mücadele tarihi açısından sonuçları ne olurdu?
HTR 9	1923 yılında gerçekleşen hangi gelişme sonrasında 'Meclis Hükümeti' sisteminden 'Kabine Sistemi'ne geçilmiştir?
HTR 10	Lozan Antlaşması ile elde edilen ve Anadolu'nun Batılı devletlerin hammadde kaynağı ve pazarı olma durumundan kurtarılması amacıyla atılmış olan en önemli adım nedir?
HTR 11	Osmanlı Devleti'nin I. Dünya Savaşı'nı müttefiki Almanya ile birlikte kazanması halinde, bunun sonuçları ne olabilirdi?
HTR 12	Sakarya Meydan Muharebesi'nin kaybedilmesi halinde ortaya çıkabilecek sonuçları değerlendiriniz.
HTR 13	Temel amacı, Türkleri kademeli olarak önce Avrupa içlerinden, sonra Balkanlar'dan ve sonuçta Anadolu'dan da atıp Orta Asya'ya sürmek olan ve Avrupalılarca güdülen siyaset, hangi kavramla açıklanmaktadır?
HTR 14	Atatürk döneminde hukuk alanında yapılan devrimlerden üç tanesini yazınız?
HTR 15	Mustafa Kemal Paşa'nın, Birinci Dünya Savaşı sona ererken, Mondros Mütarekesi'nin hemen öncesinde (1918'in Ekim ayı içinde) veya İstanbul'a geldiğinde (13 Kasım 1918) Harbiye Nazırı görevine getirilmesi durumunda ortaya çıkabilecek sonuçları değerlendiriniz.
HTR 16	Atatürk döneminde sosyal alanda yapılan devrimlerden üç tanesini yazınız.
HTR 17	Takrir-i Sükûn Kanunu hangi gelişme sonrasında çıkarılmıştır?
HTR 18	Balkan Devletlerinin (Karadağ, Yunanistan, Sırbistan ve Bulgaristan) kendi aralarında ittifak yaparak 1912'in Ekim ayında Osmanlı Devleti'ne savaş açmaları ve Birinci Balkan Savaşı'nı başlatmalarının temel sebebi nedir?
HTR 19	Atatürk'ün dış politika ilkelerinden üç tanesinin adını yazınız.
HTR 20	II. Meşrutiyet döneminde (1909 yılında) yaşanan hangi gelişme Sultan II. Abdülhamit'in tahttan indirilerek yerine Sultan V. Mehmet Reşat'ın tahta çıkmasına neden olmuştur?
HTR 21	II. Dünya Savaşı yıllarında (1939–1945) yaşanan sıkıntıları atlatmak amacıyla devletin ekonomiye müdahalesini sağlamak doğrultusunda çıkarılmış olan kanunun adı nedir?
HTR 22	II. Dünya Savaşı'nın Avrupa'da son bulmasının hemen ardından 7 Temmuz 1945'te kuruluş dilekçesi verilen ve dönemin çok partili siyasi hayata geçişindeki ilk parti olan partinin adı nedir?
HTR 23	Türkiye'nin NATO üyeliği hangi iktidar döneminde gerçekleşmiştir?
HTR 24	Özellikle Makedonya bölgesinde yürüttüğü faaliyetler ile Meşrutiyet'in yeniden ilanına (II. Meşrutiyet'in ilanına) neden olan cemiyetin adı nedir?
HTR 25	Türkiye'nin 1952'de NATO'ya (Kuzey Atlantik Antlaşması Örgütü) üye olmaması durumunda, bunun Türkiye üzerindeki etkileri ne olabilirdi?

HTR 3 was replaced before the ASAG evaluation, and is hence not included here

TL1 Items

Item ID	SA item
TL1 1	olanak – tutanak kopuk – yatak Yukarıda verilen sözcük çiftlerinden hangisindeki sözcükler aynı türetim ekini taşımaktadır?
TL1 2	"İki veya daha çok şey arasında konum, biçim ve belirli bir eksene göre ölçü uygunluğu" olarak tanımlanan sözcük hangisidir?
TL1 3	"Her türlü nitelik bakımından eşit olan, aralarında fark bulunmayan; aynı" olarak tanımlanan sözcük hangisidir?
TL1 4	"Söylediklerimden hiçbir şey anlamamış olman, beni yeterince dinlemediğini gösteriyor" tümcesinin öznesi nedir?
TL1 5	Karanfil-bitki, kedi-hayvan, gömlek-giysi gibi sözcük çiftleri, hangi anlam ilişkisini örneklemektedir?
TL1 6	Türkçede "söz dinlemez olmak, hiçbir şeyden çekinmemek" anlamlarından kullanılan deyim hangisidir?
TL1 7	Dilbilgisinde nesnesi olmayan/nesne alamayan eylemlere ne ad verilir?
TL1 8	"Ağaçlar sonbahar mevsiminde yapraklarını dök-…." tümcesinde, söz konusu eylemin geçmişte tekrarlandığı ama artık söz konusu olmadığı anlatılmış olur?

Item ID	SA item
TL1 9	"büyük sanatçı" ifadesinde geçen "büyük" sıfatının karşıt anlamlısı ne olabilir?
TL1 10	Bir kişi başkalarına saygılı olduğu kadar, kendi varlığına karşı da aynı tutumu sergilemelidir. ……, özgüvenin temelidir. Yukarıdaki tümcede boş bırakılan yere gelebilecek en uygun sözcük nedir?
TL1 11	"Toplantıya birkaç kişi katıldı" tümcesindeki birkaç kişi, kaç kişi olabilir?
TL1 12	Türkçede "davranışı kaba, sert ve gönül kırıcı olan, nadan" anlamında kullanılan sözcük hangisidir?
TL1 13	"Çocuğun eli kanadı" tümcesinin edilgen biçimi nedir?
TL1 14	"Uzmanlar, algılarımızla ulaşamadığımız pek çok bilgiye, mekanizması henüz tam olarak açıklanamamış olsa da ……. sayesinde ulaşabildiğimizi söylüyor." Yukarıda tümcede boş bırakılan yere, tümcenin içerik olarak gerektirdiği sözcük ya da ifadeyi yazınız.
TL1 15	İyi, doğru, güzel gibi nitelemeler, büyük ölçüde öznel, göreceli, ölçütleri tam olarak belirlenmemiş …… kavramlardır. Yukarıda tümcede boş bırakılan yere, tümcenin içerik olarak gerektirdiği sözcük ya da ifadeyi yazınız.
TL1 16	a. Kedi bahçeye çıktı. b. Kedi dört ayaklıdır. Bu iki tümcedeki "kedi" sözcüğünün anlamları arasındaki fark nedir?
TL1 17	Sigara böreği, küpe çiçeği, ekmek ayvası bileşik sözcüklerinin, anlamsal açıdan ortak özelliği nedir?
TL1 18	Türkçede "el elin eşeğini türkü çığırarak ararmış" atasözü hangi durumlarda kullanılır?
TL1 19	"Şiir, dilin yaramaz çocuğudur" ifadesinde şiirin hangi özelliğine vurgu yapılmaktadır?
TL1 20	Öğrencilerinin tümü İzmirli olan bir sınıf için söylenen "Bu sınıftaki bazı öğrenciler İzmirlidir" tümcesi doğru mudur, yanlış mıdır? Neden?
TL1 21	Doğru, her insan için farklı olabilir. Bu nedenle tek bir doğru'dan söz edilemez. Hatta dünyadaki insan sayısı kadar "doğru" olduğu söylenebilir. Yukarıdaki ifadeye göre, "yanlış"ın tanımı nasıl yapılabilir?
TL1 22	Tavuk-yemek ve tavuk-kuş çiftlerini oluşturan sözcükler arasında, anlam ilişkisi bakımından hangi fark vardır?
TL1 23	"Ali günde 3 ekmek yiyebilir" tümcesinden hareketle, Ali'nin her gün 3 ekmek yediği çıkarımı yapılabilir mi? Neden?
TL1 24	"Nobran" sözcüğü ne anlama gelmektedir?
TL1 25	"Ömür törpüsü" deyimi ne anlama gelmektedir?

SCL Items

Item ID	SA items
SCL 1	Karbon ayakizini azaltacak bireysel düzeyde yapılabilecek 3 şey nedir açıklayınız
SCL 2	Benzinli araçlara göre elektrikli araç kullanımının iklim değişikliği etkisini azaltmadaki rolünü değerlendirin.
SCL 3	İklim değişikliğinin ruh sağlığı üzerindeki etkisini tartışın.
SCL 4	Sanayileşme ve iklim değişikliği arasında nasıl bir ilişki vardır?
SCL 5	Yenilenebilir enerji iklim değişikliği ile mücadelede nasıl bir rol oynar?
SCL 6	Karbon ayak izi nedir?
SCL 7	İklim değişikliğinin insan sağlığına etkisine yönelik 3 örnek veriniz.
SCL 8	İklim nedir?
SCL 9	Deniz ve okyanus kenarındaki şehirlerin iklim değişikliğinden nasıl etkilenebileceğini açıklayınız.
SCL 10	Ormanların tahribatlarının 2 adet nedenini ve iklim değişikliğine etkisini açıklayınız.
SCL 11	İklim değişikliği su kaynaklarının kullanılabilirliğini ve kalitesini nasıl etkiler?
SCL 12	Tarımın iklim değişikliği üzerindeki etkisini azaltmaya yönelik bir strateji önerin.
SCL 13	İklim değişikliği ve günümüzde yaşanan ekstrem havaolayları arasında nasıl bir ilişki vardır?
SCL 14	Bitkisel ve et bazlı beslenmenin iklim değişikliğine olası etkisini karşılaştırarak açıklayın.
SCL 15	Okyanus asidifikasyonu nedir?
SCL 16	İnsan kaynaklı sera gazı etkisi nedir?
SCL 17	Bilim insanları iklim değişikliği ile verileri nasıl toplamaktadır?
SCL 18	Sera gazı emisyonlarının azaltılmasında karbon fiyatlandırmasının rolünü değerlendirin.
SCL 19	İklim değişikliğinde ozon tabakası incelmesinin ilişkisini açıklayınız.
SCL 20	Et tüketiminin sera gazı emisyonlarına olan etkisini analiz ederek açıklayınız.
SCL 21	İklim değişikliğiyle mücadelede uluslararası işbirliğinin rolünü değerlendirin.

Item ID	SA items
SCL 23	Bir sera gazı olan metanın salımının en önemli mekanizmaları nelerdir açıklayınız
SCL 24	Ulusal anlaşmalardan Paris Anlaşmasının iklim değişikliği ile mücadelede etkinliğini değerlendirin.
SCL 25	Albedo etkisi nedir ve küsesel sıcaklıklara nasıl etki eder açıklayınız.

SCL 22 was replaced before the ASAG evaluation, and is hence not included here

Acknowledgements
We extend our gratitude to the students who participated in the representative samples and cognitive interviews.

Author contributions
MK & GC: formal analysis, investigation, data curation, writing—review and editing. SP, YB & SMK: software, investigation, writing—review and editing. SG, EKD & HA: conceptualization. SA, BN & NK: conceptualization, investigation, data curation, writing-original draft. TK & BA: conceptualization, software, validation, formal analysis, investigation, resources, data curation, writing, supervision, project administration, funding acquisition.

Funding
This work is supported by Tübitak, 123K835. This publication was funded by the Open Access Publication Fund of Leuphana University Lüneburg.

Data availability
Study items and evaluation data are shared in the manuscript. The college-student-level datasets without any personal information, rubrics, and the structured outputs from the Gemini API are available from the corresponding author upon reasonable request.

Declarations

Ethics approval and consent to participate
Ethical approval for this study was obtained from the Social and Human Sciences Scientific Research and Publication Ethics Board (BAYEK) at Ege University (March 2023, protocol no 2082) in accordance with the Turkish Higher Education Council Regulations. All participants were informed about the study's purpose and provided explicit consent prior to participation. All participants were college students. Informed consent for data collection and publication was obtained from all individual participants in accordance with the BAYEK regulations. Specifically, all participants had to read and confirm the following statement to participate: "*Before participating in the study, I understand the study, its purpose, the confidentiality of my information, and my responsibilities as a participant. I understand that I will not face any negative consequences if I withdraw from the research. I consent to participate in the study of my own free will and without any coercion.*"

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 31 March 2026 / Accepted: 7 August 2026
Published online: 21 August 2026

References

1. Alderson JC, Wall D. Does washback exist? Appl Linguist. 1993;14(2):115–29. https://doi.org/10.1093/applin/14.2.115.
2. Alkhuzaey S, Grasso F, Payne T, Tamma V. Text-based question difficulty prediction: a systematic review of automatic approaches. Int J Artif Intell Educ. 2023. https://doi.org/10.1007/s40593-023-00362-1.
3. American Educational Research Association, American Psychological Association, National Council on Measurement in Education. Standards for educational and psychological testing. American Psychological Association; 1999.
4. American Educational Research Association, American Psychological Association, National Council on Measurement in Education. Standards for educational and psychological testing. American Educational Research Association; 2014.
5. American Association for the Advancement of Science (AAAS). Assessment in the service of instruction. American Association for the Advancement of Science; 1990.
6. Australian Council for Educational Research. PISA 2006 item submission guidelines for science. ACER; 2004.
7. Bacon DR. Assessing learning outcomes: a comparison of multiple-choice and short-answer questions in a marketing context. J Mark Educ. 2003;25(1):31–6. https://doi.org/10.1177/0273475302250570.
8. Baldwin D, Fowles ME, Livingston SA. Guidelines for constructed-response and other performance assessments. Educational Testing Service; 2005.
9. Barkaoui K. Effects of marking method and rater experience on ESL essay scores and rater performance. Assess Educ Princ Policy Pract. 2011;18(3):279–93. https://doi.org/10.1080/0969594X.2010.526585.
10. Behizadeh N, Pang EP. Awaiting a new wave: the status of state writing assessment in the United States. Assess Writing. 2016;29:25–41. https://doi.org/10.1016/j.asw.2016.05.003.
11. Belov D, Lüdtke O, Ulitzsch E. A supervised learning approach to estimating IRT models in small samples. Br J Math Stat Psychol. 2025. https://doi.org/10.1111/bmsp.12396.
12. Boaventura D, Faria C, Guilherme E. Impact of an inquiry-based science activity about climate change on development of primary students' investigation skills and conceptual knowledge. Interdisciplinary J Environ Sci Educ. 2020;16(4):e2225. https://doi.org/10.29333/ijese/8554.

13. Boduroğlu E, Koç O, Yiğiter MS. Madde güçlüklerinin tahmin edilmesinde uzman görüşleri ve ChatGPT performansının karşılaştırılması. Disiplinlerarası Eğitim Araştırmaları Dergisi. 2023;7(15):202–10. https://doi.org/10.57135/jier.1296255.
14. Bolgova O, Ganguly P, Ikram M, Mavrych V. Evaluating large language models as graders of medical short answer questions: a comparative analysis with expert human graders. Med Educ Online. 2025. https://doi.org/10.1080/10872981.2025.2550751.
15. Brennan RL. Statistics for social science and public policy: Generalizability theory. Springer; 2001.
16. Brookhart SM, Nitko AJ. Educational assessment of students. 7th ed. Pearson Education; 2014.
17. Burrows S, Gurevych I, Stein B. The eras and trends of automatic short answer grading. Int J Artif Intell Educ. 2015;25:60–117. https://doi.org/10.1007/s40593-014-0026-8.
18. Chang L, Ginter F. Automatic short answer grading for finnish with ChatGPT, 23173–81; 2024. https://doi.org/10.1609/aaai.v38i21.30363
19. Clauser BE, Margolis MJ, Clyman SG, Ross LP. Development of automated scoring algorithms for complex performance assessments: a comparison of two approaches. Journal of Educational Measurement. 1997;34(2):141–61. https://doi.org/10.1111/j.1745-3984.1997.tb00511.x.
20. Collins D, editor. Cognitive interviewing practice. SAGE Publications Ltd; 2015. https://doi.org/10.4135/9781473910102.
21. Cüvitoğlu G, Aydın B, Atılgan H, Kışla T. (2025). Automated short answer grading from an educational assessment perspective: a systematic review. https://doi.org/10.31234/osf.io/geayp_v1
22. Dabrowska E. Language, mind and brain: Some psychological and neurological constraints on theories of grammar. Edinburgh University Press; 2019.
23. Dawson V, Carson K. Introducing argumentation about climate change socioscientific issues in a disadvantaged school. Res Sci Educ. 2020;50(3):863–83. https://doi.org/10.1007/s11165-018-9715-x.
24. Del Gobbo E, Guarino A, Cafarelli B, Grilli L. GradeAid: a framework for automatic short answers grading in educational contexts—design, implementation and evaluation. Knowl Inf Syst. 2023. https://doi.org/10.1007/s10115-023-01892-9.
25. DeMaio TJ, Rothgeb J, Hess J. Improving survey quality through pretesting. In: Proceedings of the section on survey research methods, vol. 3. American Statistical Association; 1998. p. 50–8.
26. Dikli S. An overview of automated scoring of essays. J Technol Learn Assess. 2006;5(1):1–35. https://ejournals.bc.edu/index.php/jtla/article/view/1640.
27. Diyab A, Frost RM, Fedoruk BD, Diyab A. Engineered prompts in ChatGPT for educational assessment in software engineering and computer science. Educ Sci. 2025;15(2):156.
28. Downing SM, Haladyna TM, editors. Handbook of test development. Lawrence Erlbaum Associates; 2006.
29. Drasgow F, editor. Technology and testing: improving educational and psychological measurement. 1st ed. Routledge; 2015. https://doi.org/10.4324/9781315871493.
30. Ercikan K, McCaffrey D. Optimizing implementation of artificial-intelligence-based automated scoring: an evidence centered design approach for designing assessments for AI-based scoring. J Educ Meas. 2022. https://doi.org/10.1111/jedm.12332.
31. Finn B, Arslan B, Walsh M. Applying cognitive theory to the human essay rating process. Appl Measur Educ. 2020;33(3):223–33. https://doi.org/10.1080/08957347.2020.1750405.
32. Gedik TA. Print exposure leads to individual differences in the Turkish aorist. Lang Sci. 2024;104:101632.
33. Gedik TA. Literacy at work: ultimate native language attainment. Volume 72. Walter de Gruyter GmbH & Co KG; 2026.
34. Gemini. Google Gemini [Large language model]; 2026. https://gemini.google.com
35. Georgiev P, Lei VI, Burnell R, Bai L, Gulati A, Batsaikhan BO. Gemini 1.5: unlocking multimodal understanding across millions of tokens of context. arXiv preprint arXiv:2403.05530; 2024.http://arxiv.org/abs/2403.05530
36. Gibbs WJ. An approach to designing computer-based evaluation of student constructed responses: effects on achievement and instructional time. J Comput High Educ. 1995;6(2):99–119. https://doi.org/10.1007/BF02941040.
37. Gitomer DH. Design principles for constructed-response tasks: assessing subject-matter understanding in NAEP. ETS Unpublished Research Report. Educational Testing Service; 2007.
38. Grévisse C. LLM-based automatic short answer grading in undergraduate medical education. BMC Med Educ. 2024. https://doi.org/10.1186/s12909-024-06026-5.
39. Haladyna TM. Writing test items to evaluate higher order thinking. Allyn & Bacon; 1997.
40. Haladyna TM, Rodriguez MC. Developing and validating test items. 1st ed. Routledge; 2013. https://doi.org/10.4324/9780203850381.
41. Haller S, Aldea A, Seifert C, Strisciuglio N. Survey on automated short answer grading with deep learning: from word embeddings to transformers [preprint]. arXiv. 2022. https://doi.org/10.48550/arXiv.2204.03503.
42. Hamner B, Frasco M. Metrics: evaluation metrics for machine learning (Version 0.1.4) [R package]; 2018. https://doi.org/10.32614/CRAN.package.Metrics
43. Hasanah U, Permanasari AE, Kusumawardani SS, Pribadi FS. A review of an information extraction technique approach for automatic short answer grading. In: 2016 1st international conference on information technology, information systems and electrical engineering (ICITISEE). IEEE; 2016. p. 192–6. https://doi.org/10.1109/ICITISEE.2016.7803072
44. Haudek K, Zhai X. Examining the effect of assessment construct characteristics on machine learning scoring of scientific argumentation. Int J Artif Intell Educ. 2023. https://doi.org/10.1007/s40593-023-00385-8.
45. Ho A. Artificial intelligence and educational measurement: opportunities and threats. J Educ Behav Stat. 2024;49:715–22. https://doi.org/10.3102/10769986241248771.
46. Hogan TP, Murphy G. Recommendations for preparing and scoring constructed-response items: what the experts say. Appl Meas Educ. 2007;20(4):427–41. https://doi.org/10.1080/08957340701580736.
47. Hogan TP. Psychological testing: A practical introduction. 3rd ed. John Wiley & Sons; 2013.
48. Hrich N. Artificial intelligence item analysis tool for educational assessment: case of large-scale competitive exams. Int J Inform Educ Technol. 2024. https://doi.org/10.18178/ijiet.2024.14.6.2107.
49. Hsu F, Lee H, Chang T, Sung Y. Automated estimation of item difficulty for multiple-choice tests: an application of word embedding techniques. Inf Process Manag. 2018;54:969–84. https://doi.org/10.1016/j.ipm.2018.06.007.
50. Huang J. How accurate are ESL students' holistic writing scores on large-scale assessments? A generalizability theory approach. Assess Writing. 2008;13(3):201–18. https://doi.org/10.1016/j.asw.2008.10.002.

51. Huang J, Whipple PB. Rater variability and reliability of constructed response questions in New York state high-stakes tests of English language arts and mathematics: implications for educational assessment policy. Humanit Soc Sci Commun. 2023;10:860. https://doi.org/10.1057/s41599-023-02385-4.
52. Huang J, Zhu D, Xie D, Shu T. Examining the reliability of an international Chinese proficiency standardized writing assessment: implications for assessment policy makers. Assess Writing. 2023;55(1):100693. https://doi.org/10.1016/j.asw.2023.100693.
53. Huot B. Reliability, validity, and holistic scoring: what we know and what we need to know. Coll Compos Commun. 1990;41:201–13. https://doi.org/10.2307/358160.
54. Hussein MA, Hassan H, Nassef M. Automated language essay scoring systems: a literature review. PeerJ Comput Sci. 2019;5:e208. https://doi.org/10.7717/peerj-cs.208.
55. Ince E, Kutlu A. Web-based Turkish automatic short-answer grading system. Nat Lang Process Res. 2021. https://doi.org/10.2991/nlpr.d.210212.001.
56. Johnson RL, Penny JA, Gordon B. Assessing performance: designing, scoring, and validating performance tasks. The Guilford Press; 2009.
57. Jung JY, Tyack L, von Davier M. Automated scoring of constructed-response items using artificial neural networks in international large-scale assessment. Psychol Test Assess Model. 2022;64(4):471–94. https://www.psychologie-aktuell.com/fileadmin/Redaktion/Journale/ptam_2022-4/PTAM_2022-4_5.pdf.
58. Jung J, Tyack L, von Davier M. Towards the implementation of automated scoring in international large-scale assessments: scalability and quality control. Comput Educ Artif Intell. 2025;8:100375. https://doi.org/10.1016/j.caeai.2025.100375.
59. Kara A, Kara Z, Yıldırım S. Answer-based and reference-based BERT models for automatic scoring of Turkish short answers: the decisive role of task complexity. Int J Assess Tools Educ. 2025. https://doi.org/10.21449/ijate.1687429.
60. Kastner M, Stangl B. Multiple choice and constructed response tests: do test format and scoring matter? Procedia - Social Behav Sci. 2011;12:263–73. https://doi.org/10.1016/j.sbspro.2011.02.035.
61. Kojima T, Gu SS, Reid M, Matsuo Y, Iwasawa Y. Large language models are zero-shot reasoners. Adv Neural Inf Process Syst. 2022;35:22199–213.
62. Lambert JL, Bleicher RE. Argumentation as a strategy for increasing preservice teachers' understanding of climate change, a key global socioscientific issue. Int J Educ Math Sci Technol. 2017;5(2):101–12. https://doi.org/10.18404/ijemst.21523.
63. Lewis J, Casabianca JM. Limitations of QWK in evaluating automated and human scoring systems. Educational Measurement: Issues Pract. 2026;45(1): Article e70017. https://doi.org/10.1111/emip.70017.
64. Livingston SA. Constructed-response test questions: Why we use them; how we score them (R&D Connections No. 11). Educational Testing Service; 2009. https://www.ets.org/Media/Research/pdf/RD_Connections11.pdf.
65. Liu S, Kunnan AJ. Investigating the application of automated writing evaluation to Chinese undergraduate English majors: a case study of WriteToLearn. Calico J. 2016;33(1):71–91. https://doi.org/10.1558/cj.v33i1.26380.
66. Lottridge S, Ormerod C, Jafari A. Psychometric considerations when using deep learning for automated scoring. In: Yaneva V, von Davier M, editors. Advancing natural language processing in educational assessment. Routledge; 2023. p. 15–30. https://doi.org/10.4324/9781003278658-3.
67. Lottridge S, Wood S, Shaw D. The effectiveness of machine score-ability ratings in predicting automated scoring performance. Appl Meas Educ. 2018;31(3):215–32. https://doi.org/10.1080/08957347.2018.1464452.
68. Maeda H. Field-testing multiple-choice questions with AI examinees: English grammar items. Educ Psychol Meas. 2024;85:221–44. https://doi.org/10.1177/00131644241281053.
69. Martinez ME, Bennett RE. A review of automatically scorable constructed-response item types for large-scale assessment. ETS Res Rep Ser. 1992;1992(2):i–34. https://doi.org/10.1002/j.2333-8504.1992.tb01473.x.
70. Messick S. Validity and washback in language testing. Lang Test. 1996;13(3):241–56. https://doi.org/10.1177/026553229601300302.
71. McClellan CA. Constructed-response scoring—Doing it right (R&D Connections No. 13). Educational Testing Service; 2010. https://www.ets.org/research.
72. Mehrens WA, Lehmann IJ. Measurement and evaluation in education and psychology. 2nd ed. Houghton Mifflin Company; 1991.
73. Millî Eğitim Bakanlığı. Millî Eğitim Bakanlığı 2013 yılı idare faaliyet raporu. Strateji Geliştirme Başkanlığı; 2013. https://sgb.meb.gov.tr/yayinlarimiz/yayin/16.
74. Millî Eğitim Bakanlığı. Millî Eğitim Bakanlığı 2018 yılı idare faaliyet raporu. Strateji Geliştirme Başkanlığı; 2018. https://sgb.meb.gov.tr/yayinlarimiz/yayin/24.
75. Miller MD, Linn RL, Gronlund NE. Measurement and assessment in teaching. 11th ed. Pearson; 2013.
76. Namdar B. Teaching global climate change to pre-service middle school teachers through inquiry activities. Res Sci Technological Educ. 2018;36(4):440–62. https://doi.org/10.1080/02635143.2017.1420643.
77. National Research Council. National science education standards. National Academies Press; 1996.
78. Nozawa Y. Considerations for use of constructed-response items in educational achievement tests. Annu Rep Educ Psychol Japan. 2019;58:131–48. https://doi.org/10.5926/arepj.58.131.
79. Olivos F, Liu M. ChatGPTest: opportunities and cautionary tales of utilizing AI for questionnaire pretesting. Field Methods. 2024;37:277–90. https://doi.org/10.1177/1525822x241280574.
80. Organisation for Economic Co-operation and Development (OECD). Education at a Glance 2016: OECD Indicators. Paris: OECD Publishing; 2016. https://doi.org/10.1787/eag-2016-en.
81. Organisation for Economic Co-operation and Development (OECD). Education at a Glance 2018: OECD Indicators. Paris: OECD Publishing; 2018. https://doi.org/10.1787/eag-2018-en.
82. Organisation for Economic Co-operation and Development (OECD). PISA 2021 integrated design. OECD Publishing; 2021.
83. Organisation for Economic Co-operation and Development (OECD). PISA 2009 assessment and analytical framework: Key competencies in reading, mathematics and science. OECD Publishing; 2010. https://doi.org/10.1787/9789264062658-en.
84. Osterlind SJ, Merz WR. Building a taxonomy for constructed-response test items. Educ Assess. 1994;2(2):133–47. https://doi.org/10.1207/s15326977ea0202_2.
85. Owan V, Abang K, Idika D, Etta E, Bassey B. Exploring the potential of artificial intelligence tools in educational measurement and assessment. Eurasia J Math Sci Technol Educ. 2023. https://doi.org/10.29333/ejmste/13428.

86. Page E. Project Essay Grade: PEG. In: Shermis MD, Burstein JC, editors. Automated essay scoring: A cross-disciplinary perspective. Routledge; 2003. p. 43–54. https://doi.org/10.4324/9781410606860.
87. Palermo C, Wibowo A. Automated essay evaluation at scale: Hybrid automated scoring/hand scoring in the summative assessment program. In: Shermis MD, Wilson J, editors. The Routledge International Handbook of Automated Essay Evaluation. Routledge; 2024. p. 23–39. https://doi.org/10.4324/9781003397618-3.
88. Posit Team. RStudio: Integrated development environment for R [Computer software]. Posit Software, PBC; 2025. https://www.posit.co/.
89. Powell JL, Gillespie C. Assessment: all tests are not created equally. Annual Meeting of the American Reading Forum, Sarasota, FL, United States; 1990.
90. Powers DE, Escoffery DS, Duchnowski MP. Validating automated essay scoring: a (modest) refinement of the "gold standard." Appl Meas Educ. 2015;28(2):130–42. https://doi.org/10.1080/08957347.2014.1002920.
91. R Core Team. R: a language and environment for statistical computing [Computer software]. R Foundation for Statistical Computing; 2025. https://www.R-project.org/.
92. Raczynski K, Cohen A. Appraising the scoring performance of automated essay scoring systems—some additional considerations: which essays? Which human raters? Which scores? Appl Meas Educ. 2018;31(3):233–40. https://doi.org/10.1080/08957347.2018.1464449.
93. Raczynski KR, Cohen AS, Engelhard G, Lu Z. Comparing the effectiveness of self-paced and collaborative frame-of-reference training on rater accuracy in a large-scale writing assessment. J Educ Meas. 2015;52(3):301–18. https://doi.org/10.1111/jedm.12079.
94. Razavi P, Powers S. Estimating Item Difficulty Using Large Language Models and Tree-Based Machine Learning Algorithms. ArXiv. 2025;abs/250408804. https://doi.org/10.48550/arxiv.2504.08804.
95. Revelle W. psych: procedures for psychological, psychometric, and personality research (Version 2.5.3) [R package]. Northwestern University; 2025. https://CRAN.R-project.org/package=psych.
96. Rizopoulos D. ltm: an R package for latent variable modelling and item response theory analyses. J Stat Softw. 2006;17(5):1–25. https://doi.org/10.18637/jss.v017.i05.
97. Rogers WT, Harley D. An empirical comparison of three- and four-choice items and tests: susceptibility to testwiseness and internal consistency reliability. Educ Psychol Meas. 1999;59(2):234–47. https://doi.org/10.1177/00131649921969820.
98. Ruch GM, Stoddard GD. Comparative reliabilities of five types of objective examinations. J Educ Psychol. 1925;16(2):89–103. https://doi.org/10.1037/h0072894.
99. Ruel E, Wagner WE, Gillespie BJ. Pretesting and pilot testing. In: The practice of survey research: Theory and applications. SAGE Publications; 2016. p. 101–20. https://doi.org/10.4135/9781483391700.
100. Sahu A, Bhowmick P. Feature engineering and ensemble-based approach for improving automatic short-answer grading performance. IEEE Trans Learn Technol. 2020;13:77–90. https://doi.org/10.1109/tlt.2019.2897997.
101. Schmeiser CB, Welch CJ. Test development. In: Brennan RL, editor. Educational measurement. 4th ed. American Council on Education/Praeger; 2006. p. 307–53.
102. Sebrechts MM, Bennett RE, Rock DA. Agreement between expert system and human raters' scores on complex constructed-response quantitative items. J Appl Psychol. 1991;76(6):856–62. https://doi.org/10.1037/0021-9010.76.6.856.
103. Sheatsley PB. Questionnaire construction and item writing. In: Rossi PH, Wright JD, Andersen AB, editors. Handbook of survey research. Academic Press; 1983. p. 195–230.
104. Shermis MD, Burstein JC, editors. Automated essay scoring: A cross-disciplinary perspective. Lawrence Erlbaum Associates; 2003.
105. Shermis MD, Cohen A. International and national perspectives on machine scoring. Appl Meas Educ. 2018;31(3):175–6. https://doi.org/10.1080/08957347.2018.1464453.
106. Sridharan K, Sequeira R. Artificial intelligence and medical education: application in classroom instruction and student assessment using a pharmacology & therapeutics case study. BMC Med Educ. 2024. https://doi.org/10.1186/s12909-024-05365-7.
107. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res. 2014;15(56):1929–58. https://doi.org/10.5555/2627435.2670313.
108. Stanovich KE. Cognitive processes and the reading problems of learning disabled children: evaluating the assumption of specificity. Psychological and educational perspectives on learning disabilities; 1986. p. 87–131.
109. Sudman S. Applied sampling. In: Rossi PH, Wright JD, Anderson AB, editors. Handbook of survey research. Academic Press; 1983. p. 145–94. https://doi.org/10.1016/B978-0-12-598226-9.50011-2.
110. Svihla V, Linn MC. A design-based approach to fostering understanding of global climate change. Int J Sci Educ. 2012;34(5):651–76. https://doi.org/10.1080/09500693.2011.597453.
111. Şahin MG, Yıldırım Y, Boztunc Öztürk N. Examining the achievement test development process in the educational studies. Participatory Educational Res. 2023;10(1):251–74. https://doi.org/10.17275/per.23.14.10.1.
112. Thelwall M. Is Google Gemini better than ChatGPT at evaluating research quality? J Data Inform Sci. 2025;10(2):1–5. https://doi.org/10.2478/jdis-2025-0014.
113. Turgut F. Eğitimde ölçme ve değerlendirme. 8. ed. Saydam Yayıncılık; 1992.
114. Ulitzsch E, Belov D, Lüdtke O, Robitzsch A. Using item parameter predictions for reducing calibration sample requirements-a case study based on a high-stakes admission test. J Educ Meas. 2025. https://doi.org/10.1111/jedm.12426.
115. Ventouras E, Triantis D, Tsiakas P, Stergiopoulos C. Comparison of examination methods based on multiple-choice questions and constructed-response questions using personal computers. Comput Educ. 2010;54(2):455–61. https://doi.org/10.1016/j.compedu.2009.08.028.
116. Way WD. Commentary on "Using human raters in constructed response scoring: understanding, predicting, and modifying performance." Appl Meas Educ. 2020;33(3):255–61. https://doi.org/10.1080/08957347.2020.1750408.
117. Weigle SC. Assessing writing. Cambridge University Press; 2002. https://doi.org/10.1017/CBO9780511732997.
118. Williamson DM, Bejar II, Hone AS. Mental model" comparison of automated and human scoring. J Educ Meas. 1999;36(2):158–84. https://doi.org/10.1111/j.1745-3984.1999.tb00552.x.
119. Williamson DM, Bejar II, Sax A. Automated tools for subject matter expert evaluation of automated scoring. Appl Measur Educ. 2004;17(4):323–57. https://doi.org/10.1207/s15324818ame1704_1.

120. Williamson DM, Xi X, Breyer FJ. A framework for evaluation and use of automated scoring. Educational Measurement: Issues Pract. 2012;31(1):2–13. https://doi.org/10.1111/j.1745-3992.2011.00223.x.
121. Willse JT. CTT: Classical test theory functions (Version 2.3.3) [R package]; 2018. https://doi.org/10.32614/CRAN.package.CTT
122. Wolf LF, Smith JK, Birnbaum ME. Consequence of performance, test motivation, and mentally taxing items. Appl Measur Educ. 1995;8(4):341–51. https://doi.org/10.1207/s15324818ame0804_4.
123. Wolfe EW. Human scoring with automated scoring in mind. In: Yan D, Rupp A, Foltz PW, editors. Handbook of automated scoring: Theory into practice. CRC Press; 2020. p. 49–68. https://doi.org/10.1201/9781351264808-4.
124. Wolfe EW, Wendler CLW. Why should we care about human raters? Appl Meas Educ. 2020;33(3):189–90. https://doi.org/10.1080/08957347.2020.1750407.
125. Zhai X, Yin Y, Pellegrino JW, Haudek KC, Shi L. Applying machine learning in science assessment: a systematic review. Stud Sci Educ. 2020;56(1):111–51. https://doi.org/10.1080/03057267.2020.1735757.
126. Zhao C, Huang J. The impact of the scoring system of a large-scale standardized EFL writing assessment on its score variability and reliability: implications for assessment policy makers. Stud Educ Eval. 2020;67:100911. https://doi.org/10.1016/j.stueduc.2020.100911.

Publisher's note