CHAPTER 7

Automated and Human Scoring in TIMSS

Allison Bookbinder
Lillian Tyack
Charlotte E. A. Aldrich
Lale Khorramdel
Tianzheng Mao

Introduction

The TIMSS fourth- and eighth-grade digital assessments include a wide variety of achievement item types utilizing a range of response formats. Items can be broadly classified into whether responses are given by selecting one or several options, or whether respondents create an answer by typing, drawing, or placing objects in certain ways, etc. The two types of response formats are often contrasted as selected- versus constructed-response formats (see Chapter 1).

The TIMSS assessments include three general types of selected-response items: single selection, in which students choose one of a finite number of response options; multiple selection, in which students choose more than one option from several response options; and compound selection, where students make a series of single selections to respond in a multi-part question. The answer options for selected-response items can be presented with various response inputs, such as traditional multiple-choice buttons, drop-down menus, or clickable images and words. Constructed-response items require students to provide an answer without being given an obvious set of explicit options to choose from, either by manipulating components of the item to form an answer (e.g., using a graphing or drawing tool, by sorting information, or dragging and dropping objects), entering a numerical input, or by providing a written response.

Most TIMSS digital items can be automatically scored. Selected-response items and constructed-response items that use numeric, drag-and-drop, or graphing tool inputs can be scored with straightforward scoring rules implemented as basic computations in statistical software. Other constructed-response items can be scored by complex algorithms, either with machine learning algorithms trained on responses paired with human expert scores or with artificial intelligence (AI) algorithms that directly implement the scoring guidelines for the item. While computer-based assessment broadened the range of constructed-response items that can be scored by algorithms, some constructed-response items still require human scoring because students can provide a wide range of responses that cannot yet be reliably processed using algorithms.

Accurate scoring of constructed-response items is critical to the reliability, comparability, and validity of the TIMSS assessment results within and across cycles. To ensure that constructed-response items are scored accurately in all countries, the TIMSS & PIRLS International Study Center, in cooperation with expert groups and participating countries, develops detailed scoring guides for each constructed-response item. These scoring guides provide descriptions and examples of acceptable responses for each score point value, including examples of correct and incorrect responses. The scoring guide is developed as part of instrument development. It outlines the criteria for a response to receive credit for the item and the specific codes to apply to student responses to designate correct and incorrect answers.

For automatically scored constructed-response items, the scoring guides serve as the basis for developing code to process the raw student response data collected by the TIMSS delivery platform and assign scoring codes for the student response. For items requiring human judgment to evaluate the response, human scorers must use the scoring guides to score responses to constructed-response items reliably and validly across countries.

Several validation and quality control steps ensure all responses are scored consistently according to the guides. For human-scored items, extensive training is provided to countries in two international meetings, where content experts train scorers in applying the scoring guides to student responses. The TIMSS Survey Operations Procedures units specify a procedure for efficiently organizing and implementing the human-scoring activities using an online scoring system, incorporating IEA standards and reliability procedures. The reliability of human scoring is assessed and documented within each country, over time (trend), and across countries.

This chapter describes the general approach to scoring the TIMSS mathematics and science items, including procedures to ensure the validity of the scoring procedures for all countries. The approach for developing scoring guides is described, along with methods and procedures for both automated scoring and human scoring. The second half of the chapter focuses on how these procedures were applied to TIMSS 2023.

TIMSS Scoring Guide Development

Scoring guides are developed for all TIMSS constructed-response items as part of instrument development for each assessment cycle, which involves several phases of review and a small pilot study to inform their development (see Chapter 1). TIMSS scoring guides state the criteria for a response to receive credit. In general, scoring guides for constructed-response items define the possible scores or codes for each item, including those for partial credit if item developers wish to allow distinctions between incorrect, partially correct, and fully correct responses. The scoring guides include examples of correct, incorrect, and, if applicable, partially correct student responses.

The goal is to train scorers across participating countries to ensure they apply the scoring guides consistently within each country, over time, and across countries. For human scoring, it is important to mitigate the impact of scorer biases, such as overly lenient or overly strict application of scoring guides, as much as possible. Many constructed-response items target higher-order cognitive processes and can require complex and varied responses, and scoring guides must distinguish between correct and incorrect types of responses that can be applied universally by scoring algorithms or human scorers.

To maintain a consistent measure of achievement over time, the scoring guides for constructed-response items in TIMSS and their application must remain the same across assessment cycles. That is, for constructed-response items brought forward from previous cycles to measure trends, the scoring guides must be kept the same between cycles and should be applied consistently over time.

The development of scoring guides alongside the newly developed constructed-response items for each TIMSS assessment cycle involves efforts by many individuals, including staff at the TIMSS & PIRLS International Study Center, National Research Coordinators (NRCs), and the TIMSS Science and Mathematics Item Review Committee (SMIRC). This process typically begins as part of the item writing workshop for each cycle conducted during an NRC meeting. During the item writing workshop, meeting participants draft scoring guides for each proposed constructed-response item. Criteria for correct and, when applicable, partially correct responses are described, as well as some examples of correct and incorrect responses. The purpose of having examples in the scoring guides is to show multiple ways students may create their answers and still be correct. These examples often demonstrate the “bare minimum” of a correct answer that can still receive credit (i.e., perhaps the least complete scientific or mathematical explanation sufficient to receive credit). This is especially important at the fourth grade when students are still developing and building their scientific and mathematical understandings. It is expected that they may provide semi-complete explanations in their answers. Therefore, there must be a clear delineation in the scoring guides between responses that may be scientifically or mathematically incomplete but reflect age-appropriate understanding and so are sufficient to receive credit, and incomplete responses that are not sufficient to receive credit. For every constructed-response item, any responses that do not meet all the requirements of an acceptable response must be scored as incorrect. For certain items, specific types of incorrect responses may be provided as examples, such as those representing common misconceptions that may arise for a particular topic or common “borderline” responses that may seem correct but are not sufficient to receive credit.

To improve the newly developed scoring guides before the field test administration for each cycle, a small pilot test is conducted in English-speaking countries to collect student responses and use them to improve the scoring guides, identifying areas where the guides need to be refined to be clearer and more precise. Some student responses are added to the scoring guides as examples to clarify ambiguous responses. These student responses also form the basis for each cycle’s international scoring training materials.

Scoring guides for TIMSS constructed-response items use one- or two-digit coding schemes. Items developed in TIMSS 2019 and earlier cycles use the two-digit scheme, and items developed in TIMSS 2023 and later use the one-digit scheme.

The two-digit coding scheme uses two-digit codes to denote correct, incorrect, and partially correct answers. The first digit of the code refers to the number of score points given to the response. For correct or partially correct responses, the first digit is 1 for one-point responses or 2 for two-point responses. For an incorrect response, the first digit is 7. The second digit of the score provides diagnostic information for correct and incorrect responses, such as indicating a specific method used to solve the problem or to track a common student misconception or error. An incorrect response not fitting a pre-defined incorrect code is given a 79 for "other incorrect." If no diagnostic categories are defined, all incorrect responses receive code 79.

The one-digit coding scheme was introduced for constructed-response items in TIMSS 2023 to streamline the scoring process. While the two-digit coding may be useful for secondary analyses, only the first digit is used in operational analysis and reporting. In the one-digit coding, the 0 replaces the 7 for incorrect responses. For correct or partially correct responses, 1 is used for one-point responses and 2 is used for two-point responses. For some items, code 7 can be assigned to an incorrect response with a notable misconception or to an incorrect response with some elements of a correct answer but is nonetheless wrong. These are typically used in the field test to finalize scoring guides and procedures for the main data collection.

Automated Scoring in TIMSS

Automated scoring in TIMSS refers to all processes that score item responses using either straightforward calculations such as recodings or computing new variables, or complex algorithms that faithfully implement the scoring guide. In this chapter, the two types of automated scoring in TIMSS are referred to as straightforward automated scoring (SAS) and complex automated scoring (CAS) procedures, respectively.

Straightforward Automated Scoring (SAS)

SAS consists of automatically assigning scores to students' raw responses based on a simple computation that matches the finite set of possible admissible responses to a selected-response option; or, for constructed-response items, to a correct score and, if specified in the scoring guide, the set of partially admissible responses to a partially correct score. All other non-empty responses are assigned an incorrect score. In contrast, empty responses are handled outside the automated scoring process and assigned either an omitted or incorrect response code, depending on whether the student attempted the item. Students' responses are stored as character strings for items scored using SAS, following a saving scheme defined for each response type. This response string encodes the response provided by the student (e.g., which option the student selected, the number input on a number pad, or grid coordinates associated with drawing a line or shape).

The raw data files collected from the digital assessment application may include multiple student responses for a single item in case of answer changes. They may include blank responses due to omission or erasing a previously given response. Before implementing SAS procedures, raw data files are pre-processed to prepare datasets for each country containing only the final responses recorded for each item. Unique values are assigned to different types of missing responses (e.g., not administered, omitted, erased), which may be scored differently depending on each item's scoring guide.

Responses to items requiring students to enter a numeric input with the number pad undergo pre-processing to convert text-based character responses into a numeric form (e.g., fractions converted to decimals). This involves cleaning the text-based response for any stray or unexpected characters.

Some special recoding may need to be applied for country-language combinations to ensure consistent scoring across countries. For example, some country-language combinations following a right-to-left format may require adaptations to some items during the instrument preparation process, resulting in deviations from the international data-saving scheme. Additionally, languages using number systems different from Arabic numerals require numerical responses to be translated for international scoring. Cultural differences in using decimal periods versus commas or thousandths separators are harmonized to produce a common format to ensure consistency of automated scoring results.

Short scripts and functions are developed to implement SAS based on the scoring guides and implemented for each item in close collaboration with the TIMSS mathematics and science coordinators. Since these straightforward rules for preprocessing and matching responses to a set of correct options are uniformly applicable for similar items, common scoring functions can be used across many items that use the same item response input type. However, some items require adaptations of SAS scripts if their scoring guides include specific deviations or additions, or if there are more possible variations in student responses (e.g., items asking students to draw lines or shapes on a grid).

Complex Automated Scoring (CAS)

Some items that required human scoring in the past can now be scored automatically using more complex algorithms. One example is visual-response items, where students are asked to place an object on an image in a certain location to show their understanding. These objects can be dragged and positioned freely, and scoring guides specify acceptable regions where the objects have to be placed. This response format can be scored with an algorithmic tool that preprocesses the image and correlates students' response images with a reference image. This process is more reliable than human scoring as it guarantees that the scoring guide is applied in the same way for all student responses across countries and over TIMSS cycles.

TIMSS also uses CAS methods to validate and help improve SAS scripts for items with grid-based graphical responses (e.g., plotted points, drawn lines, or shapes), and to identify "borderline" responses requiring human review. Responses may be identified as "borderline" because they closely resemble the correct answer but have additional lines or features that could make them incorrect. While these responses are rare, they pose a problem for SAS, such that human scoring is necessary or AI/ML-based CAS is required to identify the appropriate score assigned to these rare cases. However, when scored by humans, these responses can be scored inconsistently, often awarding credit when credit is not due or denying credit if a human scorer is overly strict and penalizes even small, but according to the scoring guide, negligible, deviations. Conversely, SAS based on coordinates may be too strict. Using a combination of SAS and CAS allows for identifying such borderline responses for individual review by content experts, resulting in improved scoring validity.

For other items, such as ones that allow freely written text-based responses or free-drawing responses that do not use a pre-specified set of objects the respondent can use, CAS can be directly implemented to improve scoring reliability or to monitor human scoring accuracy. Furthermore, in preparation for operational use, some text-response and free-drawing-response items are scored a second time using AI or ML methods, in addition to the human scoring still needed in the current TIMSS scoring process. Until AI or ML scoring has been fully developed and validated for these most complex responses for use in TIMSS, the intervention of human experts remains essential.

As of the 2023 TIMSS cycle, the use of AI or ML focuses on quality control, such as checking the consistency of human scoring. Items that undergo this type of scoring include those requiring students to create or manipulate drawings or diagrams and items requiring students to provide a written response, including more than one word or complex equations. Automated scoring for quality control of scoring more complex responses is applied using various AI and ML methods, including pixel matrix correlations, artificial neural networks (ANNs), and natural language processing (NLP).

The subsections below describe CAS methods used in TIMSS. The application of these methods in TIMSS 2023 is described later in this chapter.

Scoring with Pixel Matrix Correlations

One CAS method for graphical responses, such as points or lines plotted on a grid, is based on the correlation of pixels (colors represented as vectors of numbers) in the screenshot images against a reference image. TIMSS items scored with this method have a limited number of possible correct responses (e.g., one to three) across countries as defined by the scoring guide. This method is highly efficient and reliable and can also be used to validate and refine SAS rules. The method is considerably more efficient and reliable than human scorers due to the limited number of correct responses that need to be detected reliably without variation from the scoring guide and the minimal processing power required.

For the pixel matrix correlation method, reference images composed of correct responses based on the scoring guide are compared against student responses. If the correlation between the student response image and the target response image is above a carefully determined minimum threshold (e.g., 0.95 to 0.99), then the response is scored as correct. The threshold is defined in collaboration with content experts involved in item development to maximize the validity of the produced scores.

Scoring with Trained Artificial Neural Networks

TIMSS currently uses ANNs to score complex graphical responses and, with NLP, written responses for quality control procedures, such as checking the consistency of scoring across human scorers. These ANNs are trained in a supervised learning approach to responses and expert-generated score data.

ANNs mimic the human brain's functionality by receiving input, processing information, and making decisions. First, the ANN is given image or text responses and their human rater classifications as input to train the ANN to mimic how human experts score these responses. ANNs include multiple layers of neurons that can detect features in the training data to identify which image features (e.g., lines, corners, etc.) or text (e.g., common words) are associated with correct and incorrect responses, respectively. Finally, after many iterations of training ANNs on the training samples, the ANNs can provide their own classifications for the responses (O'Shea & Nash, 2015; Tyack et al., 2024).

ANNs are more flexible than pixel correlation for scoring graphical responses and can classify image-based responses consistently, even with substantial variations in possible responses across countries. However, they require considerably more processing power and take longer to run, and most importantly, they require training data that faithfully represents the scoring guide in the form of a large number of example responses that have been scored according to the guide with little or no deviations or improperly scored responses. Therefore, only graphical-response items with more than a few possible correct responses are currently scored using ANNs.

Supervised ML with ANNs is used to validate and improve SAS scripts and to identify borderline responses for human expert review. Compared to SAS or pixel correlation scoring, ANNs tend to be more flexible than SAS and more akin to human raters and are successful in identifying borderline responses.

ANNs can also be used to assign scores to written responses. To accomplish this, students' responses are first translated into a common language (i.e., English) and go through a pre-processing procedure using NLP (e.g., Hapke et al., 2019; Yaneva & von Davier, 2023) before ANN training and classification. This method is used to assess human scoring reliability in TIMSS, including within countries, across countries, and over time.

The use of AI scoring is shown to be promising in the sense that it can closely mimic human scoring, both for graphical responses (Tyack et al., 2024; von Davier et al., 2022) and for written responses (Jung et al., 2022; 2024). While image correlations and SAS are major factors in reducing the scoring burden for countries, the application of AI and ML methods for automated scoring in TIMSS 2023 was mainly limited to quality control and the identification of borderline responses. However, it is expected that these methods will play a larger role in scoring open-ended responses in the coming cycles of TIMSS.

Ensuring High-Quality Human Scoring in TIMSS

To ensure that human-scored constructed-responses items are scored reliably in all countries, TIMSS provides extensive training in the application of the scoring guides and conducts several scoring reliability studies for each TIMSS cycle.

International Scoring Training

International scoring training is conducted for each TIMSS cycle, where all NRCs (or country representatives appointed by the NRCs) are trained to score the constructed-response items requiring human scoring. At these training sessions, scoring guides are reviewed, and their application is explained based on a set of example student responses. Example responses are chosen to represent a range of response types and to demonstrate the guides as clearly as possible. Following the demonstration of scored example responses, the training participants apply the scoring guides to a different set of student responses that have not yet been scored. The scores assigned to these practice responses are then shared with the group, and any discrepancies are resolved.

Following international scoring training, national centers train their scoring staff on how to apply the scoring guides for the constructed-response items. That is, TIMSS follows a "train the trainers" model and expects trained NRCs or their representatives to apply and train the learned scoring procedures and guidelines, to ensure scoring guides are applied consistently across all constructed response data collected in TIMSS. NRCs are guided in creating national example responses and practice responses taken from student responses collected in their country.

Documenting Scoring Reliability

Because the consistent application of scoring rules to the raw responses collected on constructed-response items is essential for high-quality data, it is important to document the reliability of the scoring process. A high degree of scorer agreement is evidence that scorers have applied the scoring guides in the same way. The procedures for scoring the TIMSS constructed-response items include procedures for double scoring a subset of responses to document scoring reliability within each country (within-country reliability scoring), over time (trend reliability scoring), and across countries (cross-country reliability scoring).

The method for assessing the reliability of the scoring within each country is for two independent scorers to score a random sample of 200 responses for each constructed response item. The degree of agreement between the scores assigned by the two scorers is a measure of the reliability of the scoring process. In collecting the within-country reliability data, it is vital that the scorers independently score the items assigned to them, and each scorer does not have prior knowledge of the scores assigned by the other scorer. The within-country reliability scoring is integrated into the main scoring procedure and ongoing throughout the process.

The purpose of the trend reliability scoring is to measure the reliability of the scoring from one assessment cycle to the next (i.e., from TIMSS 2019 to TIMSS 2023). The trend reliability scoring requires scorers of each TIMSS cycle to score student responses collected in the previous cycle. The scores assigned by the original scorers in the previous cycle are then compared with those assigned to the same responses by the scorers in the current cycle.

For each assessment, student responses included in the trend reliability scoring (200 responses per item) are actual student responses to a set of trend items from the TIMSS trend assessment blocks collected during the previous cycle assessment administration in each country and benchmarking entity. These responses are provided to each participating country and benchmarking entity and scored through an online scoring system. All scorers who score the trend assessment blocks in a given cycle are required to participate in the trend reliability scoring. If all scorers are trained to score all trend items, the software divides the student responses equally among the scorers. If scorers are trained to score specific assessment blocks, NRCs can specify within the software which scorers will score particular blocks, and the software allocates the student responses accordingly. Like the within-country reliability scoring, the trend reliability scoring is integrated within the main scoring procedure.

Finally, cross-country reliability scoring indicates how consistently the scoring guides are applied from one country to the next. Student responses included in the cross-country reliability scoring are student responses to the same items used for the trend scoring reliability study, collected from the English-speaking countries during the previous cycle assessment administration. All scorers who can score student responses written in English are required to participate in the cross-country reliability scoring, and the student responses are equally divided among the participating scorers in each country. In most countries, the scoring exercise is completed immediately after all other scoring activities.

In addition to these traditional measures for the evaluation of scoring reliability, AI and ML CAS methods are subsequently used to further evaluate the accuracy and reliability of the human-assigned scores.

Automated Scoring in TIMSS 2023

To improve the scoring reliability of the TIMSS assessments and to also decrease the burden on countries, considerable efforts were made in TIMSS 2023 to reduce the number of constructed response items that required human scoring. This was done by increasing and enhancing the use of new open-response formats afforded by the transition to computer-based assessment as well as by building the capacity to score simple constructed-response items automatically. Compared to TIMSS 2019, TIMSS 2023 reduced the amount of human scoring and improved SAS procedures. In TIMSS 2023, 81% of fourth-grade items and 80% of eighth-grade items were automatically scored.

Selected-response items and constructed-response items involving numerical inputs, line graphs, bar graphs, and some graphical responses (e.g., drawn lines or shapes or plotted points) were automatically scored using SAS.

Automated scoring rules were developed based on scoring guides, and code was written in R (R Core Team, 2021) and Python (Van Rossum & Drake, 1995) programming languages. While many SAS rules used in the 2023 cycle were the same as those used in TIMSS 2019, the more completely implemented data capture functionality in the digital environment used in TIMSS 2023 enabled the development of new rules for previously human-scored items. SAS functions were written for the items based on scoring rules developed before the data collection by the Analysis Unit at the TIMSS & PIRLS International Study Center based scoring guides created by the item writers and the mathematics and science coordinators. Any new scoring procedures for trend items were validated by comparing item statistics between TIMSS 2019 and TIMSS 2023. In cases of disagreement, further analysis was conducted, and the NRCs were contacted to provide insight into any issues (see Chapter 10).

While most TIMSS 2023 items could be automatically scored using SAS functions, some items required unique scoring scripts, either because they had unique or complex scoring rules or scoring that required multiple separate items or item parts. Most of these items used the number pad input or were part of Problem Solving and Inquiry (PSI) Task items. One item type—graphing tool items—required unique scoring rules for each item and involved using AI methods to validate and develop SAS scripts and to identify borderline responses for expert review.

Scoring TIMSS 2023 Graphing Tool Items with SAS

TIMSS 2023 graphing tool items were primarily scored using SAS on the text-based coordinate response strings. Based on the scoring guides, these items were evaluated and scored using unique SAS functions developed in Python programming language (Van Rossum & Drake, 1995). The TIMSS mathematics coordinator was closely involved in the development process, reviewing the scores assigned by the functions iteratively and providing feedback to improve scoring accuracy. For the TIMSS 2023 data collection, the graphing tool items used a snap-to-grid function, limiting the students to plotting points and drawing lines in 1-grid length increments (or 0.5 grid length for one point-plotting item). These graphical response items differed in how challenging they were to score based on the task's complexity and the possible correct responses allowed by the scoring guide. More complex items with many possible correct responses required more time to develop and review the functions since additional scoring rules had to be created.

The SAS scripts checked whether the correct coordinates were included in the response string for graphing tool items requiring students to plot points on a grid. Then, the response was checked to see if any other coordinates were plotted. The response was scored as correct if the correct coordinates were plotted with no additional coordinates (depending on the scoring guide). If the response did not contain all the correct coordinates or if it also contained incorrect coordinates, the response was scored as incorrect.

For items requiring students to draw lines or shapes, SAS scripts ranged in complexity but used the same general procedure for evaluating the drawing: identify where on the grid the student drew lines, then evaluate the drawn lines based on correct responses from the scoring guides. Coordinates needed to be evaluated individually and in relation to other lines because a single line could consist of one or more segments, adding complexity to identifying whether the correct lines were drawn.

Scoring TIMSS 2023 Items with CAS Scoring Using Screenshots

TIMSS 2023 introduced the novel capability of validating the scoring of certain items using item response screenshots with AI and ML methods. For TIMSS 2023, two fourth-grade items with special drag-and-drop features were scored using screenshots with the pixel matrix correlation method. Twelve items using the graphing tool had SAS validated and refined using screenshots with the pixel correlation method or ANNs. Additionally, three items requiring students to draw their answers were human-scored and had the human scores validated using ANNs for quality control.

Screenshot Scoring with Pixel Matrix Correlations

Six TIMSS 2023 items were scored using the pixel matrix correlation method, listed in Exhibit 7.1, with their minimum correlation thresholds. Three items—ME82511, ME82608, and MQ82C03—were primarily scored using SAS on the coordinate strings, with the screenshot-based scores used for scoring function development and validation. One item, SQ81R05, received its primary score classifications from the screenshot-based scoring after additional pre-processing and special variation steps involving some tolerance areas. Two items, ME72119 and ME72181, used the method to double-score responses in the training sample for ANNs.

Exhibit 7.1: TIMSS 2023 Items Scored with the Screenshot Pixel Correlation Method

Item	Item Label	Minimum Correlation Threshold
SQ81R05A	Shadow at two times during the day - Left image	0.9995–0.9999
SQ81R05B	Shadow at two times during the day - Right image	0.9995–0.9999
ME82511	Plot the number of flowers Jack has on days 2, 3, and 4	0.99
ME82608	Plot the point after translation	0.98
MQ82C03	Numbers of people for plans to be the same price	0.99999
ME72119	Point to complete a parallelogram	0.98
ME72181	Translation of triangle on grid	0.97

The first step in the pixel matrix correlation scoring method was to identify correct reference responses for the matrix correlation comparisons. For items with a single correct response, one student response image that matched the correct response from the scoring guide was selected. For items with more than one correct response in the scoring guide, two or three additional student response images were selected.

The next step in the scoring method was computing correlations between the image pixel matrices. First, the reference image (or first reference image) was read into R and converted into a vector of pixel color values (from 0 for black to 1 for white) using the jpeg package (Urbanek, 2019). Next, each student response image was read in, yielding a second vector of pixel color values to compare against the first (reference) vector. The Pearson correlation between these vectors was calculated. For items with more than one correct response, each reference image was read and correlated with the student response images in the same manner.

After the correlations between the student response images and the reference image(s) were computed, the correlations were evaluated to determine whether the response could receive credit. A minimum correlation threshold was determined for each item manually while the screenshot scoring program was written. During this process, groups of similar responses with varying correlations (e.g., 0.9–0.99, 0.99–0.999) were individually reviewed. The minimum correlation required to receive credit was determined by computing the number of responses that would be misclassified at different levels according to the expert reviewer and selecting the threshold that yielded the fewest misclassifications. Depending on the item, the minimum correlation threshold ranged between 0.97 and 0.99999, where response images above the threshold would receive credit. If one of the correlations computed was above the minimum threshold for items with more than one correct reference image, the response would be classified as correct.

The pixel matrix correlation classifications were compared to those assigned by SAS scripts. Any response images where the two classifications did not match were individually reviewed by content experts. During the development of the coordinate-based machine scoring, feedback from the individual review was used to improve the accuracy of the scoring rules. After the coordinate-based scoring functions were completed, a review was conducted to identify any response images where the screenshots may not have matched the stored coordinates for further review.

Screenshot Scoring with Artificial Neural Networks

For TIMSS 2023, convolutional neural networks (CNNs) were used because they are highly flexible and accurate due to their ability to process information non-linearly (Krizhevsky et al., 2017). CNNs were used to score one TIMSS 2023 fourth-grade mathematics item involving a special type of drag-and-drop input. In addition, 10 items using the graphing tool had their SAS results validated and refined with this method. Like the pixel matrix correlation method, images were represented by pixel values (0 and 1) corresponding to color (black and white) when presented to the CNNs.

For CNNs to apply classifications, they must first be trained on a subset of the data. During this training phase, the models learn which classifications are associated with different response patterns. For TIMSS 2023, a batch of approximately 20,000 student responses per item was used as the training sample for the CNN model, using response screen-shot images and the corresponding codes assigned with SAS. Because training samples must have the appropriate score codes for neural networks to learn properly and yield accurate classifications on the remaining responses (Chollet, 2018), the responses were reviewed to ensure the classifications were accurate for the sample.

A combination of CNNs and individual reviews was used to check the accuracy of these responses and clean the training data of misclassifications using an iterative process (explored and found to be successful in Tyack et al., 2024). First, a small initial training sample was created: for each item, 10% of responses from each score category across each country were randomly selected (stratified random sampling), totaling about 2,000 responses. Content experts reviewed each sample according to the scoring guide. If any responses were misclassified, they were rescored with the appropriate code before training. Any “borderline” responses on the fringes of being correct were excluded from the training to maintain clarity in learning patterns.

For the initial modeling round, CNN models were trained on the roughly 2,000 randomly sampled response images. Next, the trained models were applied to the remaining 18,000 responses, and the CNN classifications were compared to their coordinate-based machine scores. Any response images that did not match were individually reviewed by content experts, and responses whose classifications were incorrect had their scores changed to match the proper score according to the scoring guide. Additionally, any responses that the CNNs predicted with less than 90–99% probability (depending on the item) of being in that score category were manually reviewed. This threshold yielded the most success in finding misclassifications while limiting the required expert review (Tyack et al., 2024).

This process was repeated across three additional sets of training samples, using both original and newly reviewed samples with revised scores, progressively refining the accuracy of the response image classifications from the first batch. The responses were put in a final training sample following this iterative cleaning process.

Once CNNs had been trained on the cleaned sample of responses from the first batch of screenshots, they could be applied to all other screenshots. For some items, only one CNN model was used to validate the image responses, while for other items that elicited more variation in responses with more likelihood of rare “borderline” responses, two to five CNN models applied classifications to new responses. When screenshots were received, they were classified by the final CNN model(s), and their classifications were compared to the coordinate-based scores. Additionally, responses, where the CNNs had less certainty (probabilities less than 90–99%) were also reviewed.

During the development of the graphing tool SAS functions, CNNs were applied to the items to inform revisions to the scoring rules. Any differences in classifications between the coordinate-based scoring functions and the CNNs were individually reviewed by the content experts, who provided feedback to improve the accuracy of the SAS rules. Once the scoring function development was completed, CNNs were used to identify “borderline” responses to be manually classified by an expert from the machine scoring team or the mathematics coordinator.

Exhibit 7.2 lists the graphing tool items validated with CNNs, including the CNN classification accuracies (not including "borderline" responses) and the percent of "borderline" responses reclassified during machine scoring following a manual review of the screenshots. Model performance on the graphing tool items achieved 99.21–99.97% classification accuracies. Lower rates of accuracy occurred with items involving more complex scoring.

Rare "borderline" responses (about 0.5–3%) existed for all but two items. These "borderline" responses resembled correct responses but tended to be missing certain features, have extraneous lines (information), or have stray marks that called into question whether students understood the item. Thus, "borderline" response scoring of the graphing tool items in TIMSS 2023 was guided by the mathematics coordinator based on principles of human scoring student responses.

Exhibit 7.2: TIMSS 2023 Graphing Tool Items Validated with Convolutional Neural Networks

Item	Item Label	Classification Accuracy (%)	Reclassified Responses (%)
ME61081A	Draw a line parallel to AB through C	99.75	1.08
ME61081B	Draw a line perpendicular to AB through D	99.21	1.38
ME61224	Draw angle MNP larger than a right angle	99.80	1.97
ME71177	Complete the shape given 3 conditions	99.80	3.33
ME71181	Rectangle with perimeter of 10 cm	99.89	0.40
ME71211	Path parallel to Mary's path	99.52	0.86
ME72119	Point to complete a parallelogram	99.94	0.24
ME72181	Translation of triangle on grid	99.97	0.22
ME81032	Draw second half of shape (symmetry)	99.96	0.36
ME81902	Draw lines to cut pizza into 6 equal portions	99.91	0.00

Human Scoring in TIMSS 2023

TIMSS 2023 items requiring students to type responses with words or equations or to create unique drawings were scored by humans. Human scoring was conducted by participating countries digitally via IEA's CodingExpert software. Online scoring reduces the scoring burden by excluding blank responses and automatically displaying only student responses needing scoring, including any responses selected for reliability scoring. For each response to be scored, scorers saw the item stem, student response, and the applicable scores. Scorers could provide a comment and/or mark the response for their scoring supervisor's review.

To ensure the quality of the TIMSS 2023 human scoring data, countries received training to apply scoring guides for human-scored constructed response items. Results of the TIMSS 2023 reliability studies for within-country, cross-country, and trend reliability scoring are reported in Chapter 10 of this publication.

TIMSS 2023 International Scoring Training

The TIMSS & PIRLS International Study Center content experts provided scoring training to TIMSS 2023 NRCs and scoring supervisors. This training focused on ensuring that scorers interpret the scoring guides consistently. The training also provided instructions on the standardized scoring procedures for TIMSS 2023. It was then the responsibility of the trained NRCs and scoring supervisors to apply what was learned in the training to train their national scoring teams.

The TIMSS 2023 scoring training for the field test and main data collection focused on objectively evaluating the content of the student response, consistently relying on the scoring guide, and attending to the correctness of the mathematics and science in the response without penalization for grammar, punctuation, or spelling. Items requiring more complex scoring were selected for training. The training began by reading each item in the training set aloud and explaining the overarching rationale to consider a response correct and give it credit according to the scoring guide. Next, example responses were presented along with a rationale for the scores given in each case. Finally, participants independently provided scores to a separate set of practice responses and shared their thoughts about how they scored each practice response for discussion with the content experts.

The scoring training for the TIMSS 2023 data collection was held during the 6th TIMSS 2023 NRC Meeting. The training included materials for 12 fourth-grade items (8 science items and 4 mathematics items) and 25 eighth-grade items (19 science items and 6 mathematics items). Some additional items were highlighted by the trainers because of specific response examples or when the item was similar to another item included in the training session but had a few specific differences in scoring it.

The scoring training for the TIMSS 2023 field test was held at the 4th TIMSS 2023 NRC Meeting. The field test scoring training included materials for 7 fourth-grade science items and 11 eighth-grade items (7 science items and 4 mathematics items).

Written Response Scoring Validation in TIMSS 2023

While AI and ML CAS methods were not used to score any written responses in TIMSS 2023 operationally, ANNs can score short written responses as well as image-based responses with a high level of consistency; thus, ongoing research at the TIMSS & PIRLS International Study Center is being done to assess the effectiveness of ANNs in validating human scoring in a multilingual context. This approach also explores how ANN agreement statistics can be used during item review. In TIMSS 2023, 16 short written response items were selected to validate human scoring with ANNs. Unlike graphical responses, non-English written responses are first translated into a common language (i.e., English) and pre-processed using NLP methods before ANN training and classification.

The general procedure for validation of written responses involved the following steps:

1. Automatic scoring of responses that contain only numbers or punctuation (as incorrect)
2. Machine translation of non-English language responses into English
3. Pre-processing of responses
4. Training ANNs and classifying responses as either correct or incorrect

Non-English language responses eligible for ANN classification were translated with Google Translate API or OpenAI's ChatGPT API (i.e., GPT-turbo-3.5). While many languages can be well-translated by Google Translate, ChatGPT is particularly useful for misspelled responses where context is important to the translation, as it can align translations based on the subject matter given as context for each response. First, all responses were translated into English using the Google Translate API, and initial ANNs were trained on a subset of the responses. This process was repeated with ChatGPT translations. Then, the classification agreements between human raters and ANN scores were compared to determine the final machine translation method for validation. While Google Translate was the default machine-translation method, some country-language groups had considerably higher agreement between the human-rater scores and the ANN scores using ChatGPT for translation. Therefore, ChatGPT was selected as the final machine-translation method for any country-language group, where the difference in the human-ANN agreement was 3% or higher when using ChatGPT compared to Google Translate. Furthermore, if Google Translate was the selected method for the country-language group but did not successfully translate a particular response into English, the ChatGPT translation was used. A translation was considered successful if most of the words in the translated response were in the English dictionary.

Next, pre-processing was applied to the multilingual responses translated into English to prepare them for ANN classification and modeling. Pre-processing included replacing punctuation with blank spaces, breaking the response text into a list of words (tokenization), converting to lowercase, and correcting misspellings using a unique spelling dictionary based on Levenshtein's (1966) distance approach and pyspellcheker (Barrus, 2019), an NLP tool available in Python. Following this step, stemming (reducing a word to its stem, such as "swimming" to "swim") was performed. Finally, the Bag-of-Words (BoW) method was used to extract common key features (words) and create a key feature matrix. This matrix was then used in ANN training, along with the human classifications, to classify responses. For all TIMSS 2023 items, ANNs were trained on 80% of the multilingual data translated into English. Then, the trained model was applied to the remaining 20% of the data for validation.

The agreement between the ANN scores and the human rater scores was computed for all countries, with averages ranging from 78% to 94%. Further investigations were conducted to determine if any country had a low ANN-human rater agreement relative to the other countries on the item. This process involved examining the translated responses (both Google Translate and ChatGPT) with classification disagreements. Any scoring concerns following the manual review of responses by content experts were noted as discussion points for the TIMSS 2023 item review. ANN-based scoring validation was accompanied in all cases by a review of any discrepancies conducted by TIMSS mathematics and science coordinators. Using a human-guided application of AI and ML methods was beneficial for item review and quality control while supporting the already very high quality of SAS and human scoring in TIMSS 2023.

Conclusion

TIMSS 2023 maintained high standards for scoring student responses and introduced innovative advances in automated scoring that improve operational efficiencies. The number of items requiring human judgment to score was reduced from previous TIMSS cycles through digital response capture and implementation of SAS and CAS procedures. Extensive review of procedures and outcomes ensured that automatically-assigned scores were accurate and reliable, maintaining comparability with previous assessment cycles. For items requiring human judgment, scorers in the participating countries participating in training and multiple reliability measures were documented and evaluated to ensure consistency within and across countries and over time. Furthermore, results of research conducted at the TIMSS & PIRLS International Study Center found that AI-based scoring of written responses is a promising measure to evaluate the reliability of human-assigned scores, indicating potential for expanded use in future cycles of TIMSS.

References

Barrus, T. (2019). pyspellchecker Manual (Python Package Version 0.4.0). https://ansegura7.github.io/NLP/support/pyspellchecker_manual.pdf

Chollet, F. (2018). *Deep learning with Python*. Manning Publications Co. https://www.manning.com/books/deep-learning-with-python

Hapke, H., Lane, H., & Howard, C. (2019). *Natural language processing in action: Understanding, analyzing, and generating text with Python*. Shelter Island, NY: Manning Publications Co.

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

Jung, J. Y., Tyack, L., & von Davier, M. (2024). Combining machine translation and automated scoring in international large-scale assessments. *Large-scale Assessments in Education, 12*(1), 10. https://doi.org/10.1186/s40536-024-00199-7

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM, 60*(6), 84–90. https://doi.org/10.1145/3065386

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady, 10*(8), 707–710.

O'Shea, K., & Nash, R. (2015). An introduction to convolutional neural networks. *arXiv*. https://arxiv.org/pdf/1511.08458.pdf

R Core Team. (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Tyack, L., Khorramdel, L., & von Davier, M. (2024). Using convolutional neural networks to automatically score eight TIMSS 2019 graphical response items. *Computers and Education: Artificial Intelligence, 6*. https://doi.org/10.1016/j.caeai.2024.100249

Urbanek, S. (2019). jpeg: Read and write JPEG images. R package version 0.1-8.1. https://CRAN.R-project.org/package=jpeg

Van Rossum, G., & Drake, F. L. (1995). *Python reference manual* (Vol. 111, pp. 1–52). Amsterdam: Centrum voor Wiskunde en Informatica.

von Davier, M., Tyack, L., & Khorramdel, L. (2022). Scoring graphical responses in TIMSS 2019 using artificial neural networks. *Educational and Psychological Measurement*. https://doi.org/10.1177/00131644221098021

Yaneva, V., & von Davier, M. (Eds.). (2023). *Advancing natural language processing in educational assessment* (1st ed.). New York: Routledge. https://doi.org/10.4324/9781003278658