ETS®

Best Practices for Constructed-Response Scoring

2021

Foreword

Best Practices for Constructed-Response Scoring describes a set of processes and practices to promote psychometric quality and guide development, evaluation and implementation of both human and automated constructed-response scoring procedures for written, spoken, performance and multimodal responses. Our goal in authoring this document is to both improve the practice of constructed-response scoring at ETS and help establish norms of practice in AI scoring for the broader testing field.

The best practices described in this document are inclusive of reviews and feedback received from internal and external advisory panels and experts in the fields of testing and psychometrics, across academia and other testing organizations, over the course of several years. The foundation for *Best Practices for Constructed-Response Scoring* is built on measurement industry standards including Standards for Psychological Testing and Validity and Automated Scoring in Technology and Testing: Improving Educational and Psychological Measurement.

The best practices described help fill a void in the field by clearly and thoroughly documenting the critical process of scoring constructed-response tasks and subjecting them to systematic expert review. Building upon ETS's commitment to fairness and equity in testing and ensuring that our tests and products are of the highest quality, the *Best Practices for Constructed-Response Scoring* supplements and complements ETS's library of existing standards, processes and guidelines, as well as those of the broader field of educational assessment.

I am pleased to share the first edition of *Best Practices for Constructed-Response Scoring*.

Ida Lawrence

Senior Vice President, Research & Development

Educational Testing Service

Acknowledgements

The *Best Practices for Constructed Response Scoring* are the result of persistence and dedication by the following ETS leads and contributors:

Leads

Daniel McCaffrey
Jodi Casabianca
Kathryn Ricker-Pedley
René Lawless
Cathy Wendler

Contributor

Aoife Cahill
Matthew Johnson
Anastassia Loukina
Joseph Martineau
Ourania Rotou
André Rupp

Thorough, candid, and thoughtful feedback from numerous ETS and external experts informed the *Best Practices*, helping the authors to align them to the relevant standards as well as the social and educational needs they are intended to serve. The *Best Practices* do not necessarily represent the views and opinions of these individuals or their affiliated organizations, and this acknowledgement of their participation does not imply their endorsement of the published work:

ETS Leaders

Kadriye Ercikan
John Mazzeo
Randy Bennett

ETS Constructed Response Technical Advisory Committee

Anne-Marie Alexander
Douglas Baldwin
Brent Bridgeman
Jill Burstein
Ikkyu Choi
Timothy Davey
Lawrence Davis
Neil Dorans
David Escoffery
Yoko Futagi
Nancy Glazer
Matthew Johnson
William Lewis
Longjuan Liang
Jakub Novak
Gautam Puhan
Ourania Rotou
Jonathan Schmidgall
Devon Tomasulo
Edward Wolfe
Diego Zapata
Mo Zhang

External Reviewers

Derek Briggs
Li Cai
Wayne Camara
Steve Ferrara
Shelby Haberman
Suzanne Lane
Stephen Lazer
Hillary Michaels
Rochelle Michel
Corey Palermo
Barbara Plake
Stephen Sireci
John Weiner

The *Best Practices* also benefitted from a discussion of the evaluation of automated scoring of constructed responses with the ETS Visiting Panel at its meeting.

Table of Contents

Introduction

Scope and Application

Constructed-response (CR) tasks are used in a wide variety of testing contexts. The open-ended nature of these tasks offers test takers the opportunity to possibly demonstrate complex skills in the responses they construct. However, CR tasks can fail as assessment tools if the critical evidence for the target inferences from these CR tasks cannot be extracted from the responses (Mislevy, 1994). Typically, the constructed responses cannot be unambiguously judged as either correct or incorrect, as is the case with responses to selected-response items (e.g., multiple-choice or multiple-select items). Rather, the resulting responses and products must be rated as to the levels of knowledge, skills, or abilities they demonstrate. The quality of the instructions and procedures for scoring constructed responses is as important as the quality of the tasks developed to elicit the responses. The promise of constructed-response tasks to improve assessment hinges on the rating of responses to such tasks and the validity of inferences from resulting scores and their uses (Livingston, 2009; McClellan, 2010).

This document describes a set of best practices for developing, implementing, and maintaining this critical process of scoring CR tasks. It is intended for individuals who are responsible for the quality of assessments that involve CR scoring and scores. The practices described in this document (henceforth, called the *Best Practices*) address both the use of human raters and automated scoring systems as part of the scoring process, and they cover the scoring of written, spoken, performance, or multimodal responses. The *Best Practices* are not meant to cover CR tasks that can be scored by a set of rules for evaluating responses without any uncertainty on the score (e.g., a single word response for which only one word is correct).[1] The *Best Practices* are designed to cover the scoring of CR tasks used in assessments that contribute to consequential decisions as well as for other summative and formative purposes across diverse domains, populations, and stakeholder groups. The CR tasks may be used for creating and reporting (numerical) test scores, diagnostic feedback, proficiency levels, or other types of information obtained from a test, with the intention of making decisions about test takers' knowledge, skills, and abilities.

The *Best Practices* are not designed to act as an independent guide. Instead, they serve as a supplement to the *Guidelines for Constructed-Response and Other Performance Assessments* (Baldwin, Fowles, & Livingston, 2005), *ETS Standards for Quality and Fairness* (Educational Testing Service, 2014), *ETS Guidelines for Fair Tests and Communications* (Educational Testing Service, 2015), and the *ETS Guidelines for Developing Fair Tests and Communications* (Educational Testing Service, 2021). It is also important to become familiar with those standards. The Best Practices described in this document are relevant because they are intended to support the standards and guidelines laid out on those documents. In addition, the *Best Practices* reflect and support other existing professional standards, such as the *Standards for Educational and Psychological Testing* (American Educational Research Association, the American Psychological Association, & the National Council on Measurement in Education, 2014); ITC Guidelines on Test Use (International Test Commission, 2013); the *Operational Best Practices for Statewide Large-Scale Assessment Programs* (Council of Chief State School Officers, & Association of Test Publishers,

[1] The *Best Practices* also do not cover computer-scored, complex selected-response items.

2013); and the International Organization for Standardization's Standards Catalogue (International Organization for Standardization, 2017). As with the other standards, the application of the *Best Practices* depends on professional judgment and should not be considered simply as a checklist. Reasonable effort should be made to conduct CR scoring in accordance with the Best Practices described here while recognizing that certain practices may not be relevant or technically feasible for all tests or uses of CR tasks.

It is also recognized that as part of collaborations with external clients, the developer of an assessment does not always control all aspects of the CR scoring process. When possible, adherence to the Best Practices should be part of collaborative agreements and the ramifications of not complying with these practices should be made clear to clients.

Foundational Principles

Several foundational principles underlie the Best Practices included in this document. While these principles fall outside of the practices themselves, they must be acknowledged and attended to as the practices are implemented.

- The rationale for the procedures used to score CR tasks needs to follow consistent general criteria across different use contexts (e.g., a test of skills versus content), but also needs to include considerations that are unique to each test and test use. Appropriate theoretical and/or empirical evidence needs to be provided to support the rationale, including the claims made for each intended use of the scores and information generated from CR tasks.
- There is generally no single acceptable threshold for the evaluation criteria and statistics that are used to evaluate the accuracy and effectiveness of CR scoring procedures. What constitutes an appropriate threshold depends on the design of the assessment; the underlying construct being measured; how the CR scores are applied; the intended inferences and actions based on the assessment results; and the unintended and negative consequences that could result from the inferences and actions. Justifications for what constitutes appropriate thresholds across all phases of CR scoring are needed.
- The concept of fairness remains a critical aspect of evaluation for CR scoring, no matter the actual scoring mechanism. The scoring procedures should neither advantage nor disadvantage any test taker or group of test takers. Additionally, as appropriate, analyses should take subgroup performance into account. However, no single set of subgroups can be defined for all assessments. Subgroups should be based on definitions relevant to the concerns of the administration (e.g., country or test center), or other substantive concerns (e.g., different instructional contexts, demographic characteristics of test takers).
- The *Best Practices* do not recommend specific methods or approaches that should be applied universally. The stakes associated with the reported scores, the degree to which the CR score contributes to the overall score, and financial or other logistical considerations all contribute to key choices related to training, qualifying, and monitoring raters as well as implementing and monitoring automated scoring. Some methods and approaches will be commonly used but the approach taken for any

specific testing program should always be tailored to the specific context of the test and its uses.

- Evaluation of any scoring procedure for any CR tasks in any setting needs to consider the impact of implementing scoring procedures on the entire testing process. This includes possible changes to the nature of the construct; the reliability of the information produced; the meaningfulness and usefulness of the reported information; the feasibility of implementation for a particular testing program; the requirements of operational maintenance; and the degree to which the scoring procedures support the validity argument underpinning both the task score and the overall score for the intended purpose.
- Finally, the criteria used to evaluate CR scoring procedures may differ depending on whether the procedure is a new one or a modification of an existing one.

1. A Framework for Establishing Validity Evidence for Constructed-Response Scores

The CR scoring practices constitute a set of activities that can be followed to produce scores from constructed responses that are better able to support the claims made by an assessment and provide evidence of the validity of those inferences than would otherwise be possible without engaging in those practices. The practices are built on the framework for inferences for constructed-response scores presented in this section. This framework was informed by multiple sources such as the chapter on validity in the *Standards for Psychological Testing* (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, 2014) and the book chapter *Validity and Automated Scoring* in *Technology and Testing: Improving Educational and Psychological Measurement* (Bennett & Zhang, 2016).

The central tenet of the framework for establishing validity evidence is that the evidence starts with the implementation of procedures meant to create assessments and resulting scores that support an assessment's claims about test takers' knowledge, skills, and abilities. Documentation of these procedures and the results of empirical studies (both qualitative and quantitative) round out the evidence. This cycle of "do-document-evaluate" operationalizes the collection of the various sources of validity evidence (Kane, 1992). For CR tasks, this do-document-evaluate cycle of evidence generation must be repeated for each of four elements that lead to scores: 1) the stimulus that elicits the test taker's constructed response; 2) the environment in which the stimulus is presented and which captures the test taker's response; 3) the rating of the constructed response; and 4) the process for creating the reported score or scores from that response.

Much of the validity evidence for scores from any type of assessment, including scores from constructed-response tasks, comes from the assessment development process and operational testing procedures. This includes what the *Standards* refer to as evidence, based on test content, response processes, and internal structure. Compared to selected-response items, constructed-response tasks require additional practices to collect this evidence through the task development, operational processes, and practices that are the focus of this document. Evidence based on other sources (e.g., convergent and discriminant or criterion relationships) is rarer, but can take on special importance with constructed-response scoring.

1.1 Task and Test Design and Administration

The assessment tasks need to be designed to elicit responses that demonstrate the knowledge, skills, and abilities of the construct of interest. In addition, the responses they elicit must provide evidence of levels of the targeted knowledge, skills, and abilities of the test takers. The development of tasks that achieve the desired goal can be greatly enhanced by following a principled approach to task and test development. Such approaches suggest processes for creating tasks that are meant to ensure the desired functioning of the tasks, and they naturally consider the complexities involved in developing the CR tasks and scoring the constructed responses. These frameworks for development fully consider all aspects of the assessment process, including the development of the construct/domain definition and purpose(s) of the assessment; the conceptual models that underlie the framework such as specifications for the evidence needed to support claims about individuals; and the processes involved in facilitating the assessment. The Evidence-Centered Design framework of Mislevy, Almond, and Lukas (2003) is a well-known example of a principled test and task development framework.

1.2 Rating the Responses

Constructed-response tasks can be scored by human raters and/or by computer algorithms. We first discuss the framework for human ratings and then turn to automated scoring by computer.

1.2.1 Human Rating

Constructed-response tasks are designed to elicit performances from test takers that provide the users of the assessment with evidence for making inferences about the knowledge, skills, and abilities of interest. However, the elicited performances or responses are not directly useful for a test user to make those inferences. They must first be evaluated in order for test users to make those inferences about the test takers. Typically, the evaluation yields a numerical rating in which the numerical values are meant to order the quality of the performance in terms of what it demonstrates about the targeted knowledge, skills, or abilities. Whether the resulting ratings support inferences about the targeted abilities depends on the process for making those ratings and how well those methods are followed when each response is evaluated.

When human raters provide these ratings, the process for rating generally entails the following steps:

- the development of a rubric which describes how elements of the performance relate to the targeted knowledge, skills, and abilities and how to assign numeric values based on those elements in a given performance;
- the development of training materials including task-specific scoring notes and annotated sample responses to train raters on how to apply the rubric as intended by the test developer;
- the development of a process for ensuring that a trained rater can apply the rubric as intended;
- the training and testing of raters to ensure they are qualified to score;
- the evaluation of performances by raters;
- the analysis of data from a rating quality control process and subsequent actions taken to manage rater quality;
- the collection and analysis of summative data on the performance of raters, and
- adjustments to the training process and materials to fine-tune rater performances for future test administrations.

With constructed-response items, the rating of the response is as essential to the production of scores as is the actual creation of the tasks and the test design. Human raters' personal experiences and beliefs can lead to biases in human ratings with the potential to degrade scores. Quality training can remove these biases (Fahim & Bijani, 2011; Weigle, 1994). Evaluations are needed to show such biases are not contributing to ratings, and monitoring is needed to ensure they do not re-emerge over time. Consequently, the development and monitoring of the human-rating process should be integral to task development and guided by a principled design framework followed for the drafting of the tasks and designing of the test.

These steps are often conducted iteratively through a combination of pilot and field testing and then based on operational results. The validity of the inferences based on the resulting ratings requires each step to function as intended. Raters cannot provide meaningful ratings if the rubric does not accurately identify the elements of the performance that represent the targeted knowledge, skills, and abilities and if the numeric values do not properly rank the performances. Moreover, raters cannot produce useful ratings if the rubrics do not communicate the decision rules properly or if the training materials and process do not train the raters to apply the rubric as intended.

The do-document-evaluate process for generating validity evidence for the human rating requires a set of best practices that explicate the actions to be taken at each step. This process includes the information to be documented; the evaluation procedures for each step; the final results that are all part of the development and operational practices of the assessment; and the creation of scores.

1.2.2 Computer-Generated (or Automated) Scores

Computer algorithms can also rate constructed-response performances. The resulting scores are often called automated scores. There are two steps to the automated scoring process:

- A computer algorithm extracts information from a digital representation of the constructed response; and
- A second algorithm takes the information extracted from the constructed response as input and assigns it a score.

The validity evidence for inferences drawn from the scores that are derived from these two steps must demonstrate that these scores can be considered as indicators of the knowledge, skills, and abilities that the assessment is meant to assess. The evidence will follow the do-document-evaluate cycle of developing, documenting, and evaluating the outputs from both steps of the process.

In the first step, information is extracted from the response to use as inputs in the second step. We use the term “feature” to refer to those inputs.[2] These features may or may not be developed to evaluate identified elements of the construct. For example, the inputs of the ETS

[2] Sometimes the term “feature” refers only to variables developed to evaluate identified elements of the construct of interest such as fluency of speech or the complexity of the vocabulary. Here the term is more inclusive, referencing any variable extracted from the constructed response; those that were and were not developed with such explicit ties to elements of the construct.

e-rater® engine algorithms used for scoring written constructed responses are features that include numeric assessments of usage, grammatical and mechanical accuracy, and the sophistication of the vocabulary, among other elements of the writing construct. Conversely, features generated by other algorithms may consist of indicators for the existence of specific character, letter, or word sequences in a constructed response (e.g., binary variables equaling 1 if the sequence is in the text and zero otherwise) and these features have no direct link to explicit elements of the construct.

The first-step algorithm that performs the extraction of information can be complex and can involve many components. For example, to evaluate the content of a spoken response, the algorithm might start by mapping a digital audio signal onto a digital textual representation ("automated speech recognition"), followed by decomposing the text into its grammatical parts, and then extracting the final features. The derivation of the features might also involve multiple steps. The validity evidence from the first step might begin with documentation of the algorithm used for extracting the texts to the automated scoring model. In particular, the documentation could include information that demonstrates how the output of the algorithm relates to the construct of interest and evidence that the algorithm functions as intended. The evidence might also include the results of empirical analyses that demonstrate the relevance of the resulting features to the construct of interest. The empirical examples are likely to include a comparison of the extracted texts to human annotations or human ratings of specific aspects of the responses.

Human evaluations and human annotations of the responses may be used in developing and testing the algorithms, which generate features that measure specific aspects of the construct to be assessed, as the interpretable linguistic features of the e-rater® engine. Similarly, empirical evaluations used to support claims about the validity of the interpretation of feature scores can include comparisons between feature scores and human ratings of specific aspects of the responses (e.g., its grammar) or human annotated data where experts have created labels that capture the aspects of the construct the feature is intended to measure.

The second step of the automated scoring process consists of the development of an algorithm to assign scores using the features generated in the first step. This second step is commonly built to predict a numerical criterion by solving a mathematical optimization problem to yield the closest predictions of the criterion on a training sample of data. In current automated scoring procedures, the criterion is almost always a human rating of the response. The accuracy of the resulting predictions typically serves as a key piece of evidence for the validity of automated scores; however, predictive accuracy is not sufficient. The evidence should again also include detailed documentation of the algorithm and evidence that the algorithm is implemented and functioning as intended. Evaluations showing that the resulting scores have the expected relationships with both construct-relevant and construct-irrelevant variables can greatly enhance the evidence for the validity of the automated scores that result from this two-stage process.

Further, just as bias in human ratings has the potential to degrade scores, there is a risk for bias in automated scoring. Not only can the human ratings upon which automated scoring models are built carry bias with them into the models, but so can other factors in model building such as determining feature weights or other activities. Evaluations of automated scoring models must ensure that models are fair across key groups. As stated by the Council of Europe, private groups that create algorithmic systems (such as automated scoring engines and models) have a responsibility to "seek to ensure that the design, development and ongoing deployment of their algorithmic systems do not have direct or indirect discriminatory effects on individuals or groups that are affected by these systems, including on those who have special needs or disabilities or who may face structural inequalities in their access to human rights" (Council of Europe, 2020, p. 23).

1.2.3 Human Ratings and Validity Evidence for Automated Scores

Human ratings play a key role in creating validity evidence of automated scores. As noted above, most of the validity evidence for scores derives from the development process and operational data. For automated scores this means that, by far, the most important sources of evidence for their relevance to the construct of interest are how closely they correspond with human ratings for the same constructed responses. It is often the only evidence provided. Consequently, the validity evidence for automated scores can often be no stronger than the evidence for the human ratings. Alternative sources of evidence for the automated scores are possible, but even when such evidence exists, comparisons of automated ratings to human scores remain a central element of the evidence for the automated scores. Moreover, computer algorithms trained to predict human ratings will reflect any biases in the human raters (Wind, Wolfe, Engelhard, Foltz, & Rosenstein, 2018). Thus, for most testing programs, it is critical for there to be strong evidence for the human ratings before developing automated scoring methods.

In this context of using human ratings for developing validity evidence for automated scores, Bennett and Zhang (2016) adapt Bejar's (2012) concept of a first-order validity argument, which requires stipulating and gathering evidence to support the *meaning of the outcome* (i.e., the human rating), in addition to stipulating and gathering evidence about the relationship between the predictor and outcome. Bennett and Zhang (2016, pp. 152-153) identify the following types of evidence as necessary when using human ratings as a criterion:

- evidence that the processes in which test takers engage align with the construct definition;
- evidence that the scoring rubric fully captures the construct definition;
- evidence that the processes in which raters engage align with the rubric;
- evidence that raters highly agree, with respect to group-level measures and individual-level indices (based on multiple measures);
- evidence that raters accurately score atypical responses;
- evidence that human ratings on one task predict performance on other tasks from the same task universe (ratings across tasks must be correlated);
- evidence that human ratings are related to other indicators; and,

- evidence that the characteristics listed above hold across important population groups.

This evidence closely aligns with the evidence discussed above in the do-document-evaluate process for generating validity evidence for the human ratings and which the best practices for human ratings (presented below) aim to produce.

However, even strong evidence for the validity of the human ratings is not sufficient for concluding that automated scores support the desired inferences, even if the automated scores correspond strongly with the human ratings. A high correlation does not mean the two variables measure a common construct. For instance, many measures of different constructs often have a strong correspondence (e.g., reliable reading and math scores often have a high correlation). Similarly, even if the automated scores and human ratings are highly correlated, their differences could be associated with attributes of the test takers in ways that could distort inferences about those test takers.

Discrepancies between automated scores and human ratings are not parallel to the discrepancies between two human ratings. Differences between two human raters indicate idiosyncratic judgments of individual human ratings.[3] If human raters are randomly assigned to the constructed responses they rate, then the differences between raters cannot be correlated with characteristics of test takers except by random chance. However, differences between automated scores and human ratings are composed of the same idiosyncratic judgments of the humans and, potentially, systematic differences between human ratings and automated scores. These differences could make inferences based on automated scores invalid, even if the evidence supports the inferences from human ratings. Thus, additional evidence that supports the automated scores can be valuable. This evidence could include results from an evaluation of the differences between human rating and the automated scores. It could also include evidence on the relevance of the features for assessing the construct of interest or correlation between the automated scores and other relevant criteria.

Evidence about automated scores should include more than the accuracy of the prediction of human ratings. For example, additional explorations of the inner workings of the prediction algorithm that indicate what information the algorithm uses to assign scores could provide evidence that scores are related to the constructs of interest so that applying them broadly is supported. Also, checks that the algorithm does not assign high scores to nonsensical responses can provide another source of evidence in support of the validity of the claims. Beyond these examples, testing programs using automated scoring should pursue other explorations to further support claims about the validity of automated score use.

The important point is that the validity evidence for automated scores typically will begin with the evidence for the validity of inferences from human ratings. Human ratings are integral to automated scoring. It would be difficult to create an automated scoring system that supports

[3] Individual raters might be prone to making systematic errors across responses, but these errors are still idiosyncratic to the rater. As a group, all raters can make systematic errors. However, if all raters make the same error, then this error would be common to any two raters and not result in differences in their scores. Thus, differences between human raters correspond to idiosyncratic judgments of the individual raters.

the claims of an assessment for which validity evidence based on human ratings is lacking. However, the accuracy of the system to predict human ratings or the close correspondence between human ratings and automated scores should be seen as only the start of the evidence. Additional sources of evidence should be sought, as they can greatly bolster the support for the use of automated scores.

2. Case Studies

We use two case studies as running examples throughout Sections 3 to 7 to demonstrate the constructed response best practices. These examples give context and provide clear descriptions, clarifying practices in more practical terms. One example is based on a writing task and uses a scoring system that relies on an automated scoring engine that processes written responses to assess writing quality. Another example is based on a social studies test that relies on an automated scoring engine that scores content-specific features of written short answer responses. Details of these running examples are in Table 1 and Table 2. Note that the scenarios and the statistics provided throughout these case studies are fictitious and do not represent results for a real testing program evaluation.

Table 1

Case Study 1: Summary of Writing Task Example

Goal of the Testing Program	General achievement test for admission to professional school with a specific evaluation of writing. The program was evaluating the inclusion of an additional feature in the automated scoring engine and scoring model.
Target Population	Adults (18+) applying for professional school
Test Structure	Selected-response section and CR section (2 essay tasks, each scored on a 1- to 5-point scale)
Reported Scores	Scores reported by test section
Intended Score Usage	Selection
Stakes (Low, Medium, High)	High stakes for the individual.
Frequency of Testing/Scoring	Tests were administered continuously. The scoring of the constructed responses was ongoing (on a daily basis).
Testing Volume	Approximately 10,000 test takers a month
CR Scoring Design	Average of human (H) and machine (M) scores. Machine scores were unrounded in score computation. Discrepancies between human and machine scores were resolved by obtaining additional human ratings.
Section Score Computation (Weighting)	Equal weighting between tasks (each task was worth 50% of the section score). CR items contributed 100% of the writing section score.
Total Score Computation	N/A

Human Scoring Design	Raters were required to have professional school degrees. Scoring occurred five days per week, year-round. Raters were scheduled for scoring sessions based on their availability. Raters trained and scored remotely using an online score capture system. They were monitored and mentored online and by phone by Scoring Leaders. All responses were assigned at least one human rating, a randomly selected agreement sample of 5% of the responses received a second human rating for quality control purposes.
Content validity evidence	Features were designed to capture specific elements of writing that research has shown to be relevant for human judgments on the quality of writing. Research supports this use for at least a subset of features.
Engine	e-rater®, a well-established engine for evaluating writing with significant historical use, including for this assessment (Attali & Burstein, 2006). For more information on e-rater in layman's language and additional links to valuable resources, visit: https://www.ets.org/erater/about
Model type	A generic scoring model was applied for all prompts (i.e., it is not prompt-specific and is used across all prompts)
Statistical Modeling Method	Non-negative least squares regression model predicting the human rating from 10 e-rater features.
Model-building variable	First human rating

Feature Descriptions	Existing features • Grammar (e.g., subject–verb agreement) • Usage (e.g., *then* versus *than*) • Mechanics (spelling and capitalization) • Style (e.g., repetitive phrases and passive voice) • Organization (e.g., thesis statement, main points, supporting details, conclusions) • Development (e.g., main points precede details) • Positive features: o Correct preposition usage (the probability of using the correct preposition in a phrase) and good collocation use (i.e., collocations occur when two adjacent words appear together more often in language use than other pairs of words, such as the pairing of tall trees and high mountains as opposed to high trees and tall mountains) o Sentence variety (i.e., the ability to use correct phrasing and a variety of grammatical structures) • Lexical complexity with average word length (i.e., the use of vocabulary with different counts of letters) • Lexical complexity with sophistication of word choice (i.e., the use of sophisticated vocabulary) Feature to be added: • Discourse Coherence, which measures the coherence of an argument by tracking use of related words throughout an essay

Table 2

Case Study 2: Summary of Social Studies Test Example

Goal of Testing Program	Interim achievement test for social studies (K-12). The goal was to introduce automated scoring for a new set of CR prompts.
Target Population	K-12 Students
Test Structure	Selected-response section and CR section (5 tasks, short answer/content based, each scored on a 0 - 2 or 0 - 1 scale)
Reported Scores	Overall scale scores and performance levels for overall and domain. CR scores were not reported separately. Students received their scores and school districts received data for their students.
Intended Score Usage	Reported to teachers as feedback on their students as part of an interim assessment program.
Stakes (L, M, H)	Low stakes for students and educators.

Frequency of Testing/Scoring	Tests were administered once a year. The scoring of constructed responses was performed after the administration over the course of three weeks.
Testing Volume	Approximately 75,000 test takers
CR Scoring Design	Machine scoring only; human scores were collected during field testing.
Section Score Computation (Weighting)	N/A
Total Score Computation	Sum of the scores of CR and selected-response tasks. The sum score was scaled using the test characteristic curve from an item response theory (IRT) model fitted to the selected-response item scores and the automated scores from the CR tasks (using rounded scores). CR tasks contributed 40% to the total score.
Human Scoring Design (field test)	Raters were required to have a college degree, preferably in a relevant field. Once a year, scoring occurred daily over a three-week period following field test data collection. Raters were scheduled for scoring sessions based on their availability. Raters trained and scored remotely using an online score capture system. They were monitored and mentored online and by phone by the Scoring Leaders. All field test responses were scored by two raters.
Construct validity evidence	Scoring model inputs were generic linguistic features that captured the structure of the wording of the responses without any specific references to content. An example would be a feature based on two-word *n*-grams – the set of all sequential two-word pairs contained in the response. For example, in the start of a sentence, "The Declaration of Independence was signed in Philadelphia…" the two-word *n*-grams are *the Declaration*, *Declaration of*, *of Independence*, *Independence was*, *was signed*, *signed in*, *in Philadelphia*, … Features were derived from these. Features based on other linguistic structures were also used but, again, they were not specific to the content by design. These features would have been sensitive to content as it was presented in the wording of the responses and judged by humans, but they were not explicitly developed to use information about the content in the calculation of the scores. There were a small number of features which equal 0 - 1 indicator variables that are equal to 1 if a keyword (e.g., Declaration of Independence) was in the response and 0 otherwise. The keywords were directly related to the content that should have been included in responses that indicated knowledge of the item content.

Engine	c-rater™. c-rater is a relatively new application at ETS. For more information on c-rater's capabilities, visit: https://www.ets.org/accelerate/ai-portfolio/c-rater/
Model type	Prompt-specific
Statistical Modeling Method	Support Vector Regression
Model-building variable	First Human Score
Feature Descriptions	Features derived from: • Sequences of characters (character *n*-grams). Account for variations in responses due to spelling errors and morphological variants (e.g., look, looks, looked, looking) • Sequences of words (word *n*-grams). Account for certain key words and phrases required at each score point. • Syntactic relations. Account for key relationships in the response at each score point. • Length (where appropriate). Accounts for amount of detail and elaboration in the response. • Indicator variables that equal 1 if a keyword is in the response and 0 otherwise.

3. Principles/Standards and Evidence for Constructed Response (CR) Task and Test Development

The validity, reliability, and fairness of the scores produced from CR tasks are dependent upon the quality of those tasks. When embedded within a principled design framework, carefully specifying the expected functioning and purpose of the task contributes to the development of tasks that yield observable evidence of the construct. Conversely, poor quality CR tasks developed without careful consideration will likely result in test-taker responses that are heavily influenced by construct-irrelevant features. This, in turn, will create scoring difficulties for raters and will lead to scores that are not useful for the intended purposes.

For this reason, this section focuses on the aspects of CR task and test development specifically related to collecting validity evidence and ensuring acceptable levels of reliability and fairness. This section does not include *all* elements of the validity argument, but the elements necessary to lay a strong foundation for the scoring process. For more comprehensive guidelines refer to the *Standards* (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, 2014) or the *Guidelines for Constructed-Response and Other Performance Assessments* (Baldwin et. al., 2005), and other resources focused on validity argumentation (Kane, 1992; Mislevy, Almond, & Lukas, 2003).

3.1 Construct Relevance of Tasks and the Constructed-Response Section of the Test

When constructed-response tasks/tests are created, task relevance as well as content coverage should be considered.

Individual CR tasks and the CR section as a whole should be developed to align with the construct definition. Each task should directly relate to the construct definition so that it can be used as an indicator, at some level, of the construct, depending on the internal structure and how subdomains or subareas of the construct contribute to an overall construct definition.

The task wording should only involve content and subjects relevant to the construct and exclude extraneous wording or complexities. The expected mode of response (spoken, performance, text, etc.) and scoring (automated or human raters) should also be considered. Furthermore, there should be no aspects of the tasks that could yield differential responses from different groups of test takers.

The rubric and/or scoring guides are critical elements of the task and test, and they are discussed below in 4.1.

Evaluation and Evidence: [4]

- Conduct an independent review of the tasks (by experts) for content coverage and relevance. The review should be performed task-by-task and at the level of the test form/test.
- Conduct an independent review of the tasks for issues related to bias, sensitivity, and lack of accessibility to ensure the context and wording of the tasks exclude any aspects that could lead to unfairness.
- Conduct a task analysis that shows the task is performing as expected and is correlated with the overall construct (via the total section/test score).
- Analyze the cognitive processes used by test takers when they respond to the task to ensure that the task consistently elicits the intended responses.
- Gather documentation of the evidence with the overall assessment of the tasks/test.

3.2 Evidence Based on Relations to External Variables

Evaluate the concordance between the task/test scores and related statistics.

We rely on scores of constructed-response tasks to capture rich information about test takers' performances on the construct of interest. To ensure that CR scores capture this information, some evaluation of the relationship between the scores and other related measures of the construct or related constructs should be conducted. This is important at the CR section/test level, and also a consideration at the CR task level.

These statistics or criteria may include selected-response task/test scores, scores on other CRs from the same test, scores from external assessments, or performance ratings from an authentic environment. Evaluating the performance relative to an internal selected-response performance, to external assessment scores, or to other criteria can demonstrate whether tasks differentiate among test takers as expected.

The quality of the criteria should be commensurate with the stakes of the assessment they are being used to evaluate. That is, the external criterion scores used for evaluation should meet the same level of psychometric and validity quality requirements that the new constructed-response task/test is intended to meet.

Evaluation and Evidence:

- Estimate the correlation between external criterion scores and task/test scores. The strength of the correlation is dependent on the how much the individual CR task is meant to reflect the full construct definition and how much the score contributes to the total score.

[4] This evidence is necessary for new tasks or task types but once the testing program is established, some of these reviews may not be needed.

- Document the correlation analyses, explaining the selection of the external variables, and provide evidence that the external variable is of high quality (evidence of validity/reliability for the external variable scores).
- If an external criterion is not available, then heavier reliance must be placed on an expert review of the ratings and their relationship to the rubric and the construct.

4. Principles/Standards and Evidence for Human Rating

When human raters are used to score CR tasks, they become an essential element in ensuring the reliability of scores and validity of the interpretation and use of those scores. Thus, it is imperative that processes to ensure that constructed-response tasks are scored by humans consistently, accurately, and in ways that capture the construct-relevant evidence are in place before operational use. There are four critical phases that must be considered: rater selection, rater training, raters' demonstration of scoring competency, and rater monitoring. The CR best practices in this section are meant to ensure that human ratings appropriately reflect the established scoring guidelines set by the test developers for the CR tasks so that interpretations made with the scores are defensible. These practices assume that best practices were followed in the creation of the CR tasks and there is evidence to support the use of the tasks.

During task development and prior to implementing CR scoring with human raters, several logistical decisions must be made including determining:

- the method (e.g., analytic, holistic, etc.) that will be used for scoring responses;
- the number of score categories (e.g., points on the scale, levels of competency);
- the number of raters who will evaluate each response;
- the sample responses that are used for training, qualifying, and monitoring purposes; and
- policies and procedures for rescoring.

In addition, the required criteria (e.g., educational background, content knowledge, specialized skills) must be established that will be used to determine if an individual is eligible to apply for work as a rater. Materials used in support of rater training, qualification testing, and scoring also must be created and reviewed by experts prior to their use. ETS's *Guidelines for Constructed-Response and Other Performance Assessments* (Baldwin et. al., 2005) provide more details on these issues.

Finally, when possible, empirical evidence that supports the selection, training, and monitoring processes used with raters should be obtained. The following principles describe best practices for human scoring and the collection and documentation of supporting evidence.

4.1 Construct Relevance of Human Scoring Process and Materials

Design and implement a human scoring system that is based on the explicit link between the construct to be measured, scoring rubric, and support materials.

To build a validity argument that supports the use of a constructed-response task, there must be a clear link between the construct and the evaluation system that is used (including the scoring rubric and materials). During task development, the scoring rubric is created to outline the requirements for responses for each performance level or score point. During task development and on an ongoing basis, scoring materials are created, updated, and maintained to operationalize the definitions presented in the rubric, including notes and guidelines specific to the task.

Before using the rubric and scoring materials, materials should be evaluated to ensure that they align with the construct.

Evaluation and Evidence:

- Document the process used to write the rubric and the link between the rubric and the construct definition.
- Conduct an independent review of the scoring materials for content coverage and relevance.
- Conduct an independent review of the scoring materials to ensure the context and wording of the materials exclude any aspects that would lead to unfairness.
- Analyze the cognitive response processes involved when a rater evaluates a performance to ensure that the use of the rubric and other scoring materials consistently yield the intended decision-making processes. This should be done as part of the initial collection of validity evidence. This may involve think-alouds, cognitive interviews, or surveys.

Case Study 1
The rubric was created to provide raters with a holistic approach to evaluating the quality of written responses. Below is a high-level description of the rubric for this task, which was aligned with the intended construct. More detailed scoring guidelines were included with the full rubric available to raters. During rubric development, think alouds, and cognitive interviews of raters were conducted. A committee of writing experts was formed, representing the academic institutions that used the assessment. They reviewed and approved the rubric.

5 - Presents a well-articulated response that conveys meaning clearly.
4 - Presents a generally well-developed response that somewhat conveys meaning.
3 - Obviously flawed but demonstrates competence in conveying meaning.
2 - Demonstrates serious weaknesses in writing.
1 - Demonstrates fundamental deficiencies in writing.

Case Study 2
There was a general rubric to score each task type, which was accompanied by prompt-specific scoring notes related to the prompt's content. For example, the general scoring rubric for Prompt 1 is provided below. A group of content experts from the *Association of Teachers of Social Studies* reviewed and approved the rubrics.

2 - Accurately presents two pieces of evidence.
1 - Accurately presents one piece of evidence.
0 - Does not accurately present any evidence.

4.2 Exemplar Selection and Scoring Process

Beyond the application of the scoring rubric developed to score a task, samples of responses, or exemplars, provide raters with the opportunity to review responses scored by experts to assist them with the scoring process. Exemplars are reviewed by scoring and content experts who agree upon a rating or consensus score, which reflect that they are based on agreement between multiple raters. Exemplars provide more concrete guidance to raters about how to apply the rubric by serving as examples of responses that merit each of the score points.

Exemplars are a critical piece of the scoring process and gathering of validity evidence. Not only are exemplars used to train and provide guidance to raters, but they are also used for quality assurance evaluation throughout the scoring process (via qualification and monitoring processes, described later in this section). They are the critical link between the scoring rubric, the specific task, and the interpretation of how the rubric is applied to real responses.

4.2.1 Principled Selection of Exemplars

Support rater training and scoring by selecting exemplars that represent the variety of possible response types, scores levels, and test-taker characteristics.

It is important that the exemplars not only cover the breadth of score points but also represent the various types of approaches and styles of responses at each of the score points. Spoken responses with many different accents, tone, or speed; written responses with various types of grammatical structure, vocabulary, or length; and responses that take different approaches are all examples of the wide variety of responses that should be reflected in a set of exemplars. Note that there are many different types of constructed-response tasks across the spectrum with a varying need for sample responses—not all of them will depend on exemplars to the same extent. However, for those tasks that necessitate them, as many responses as possible should be available for selection for this effort to be most effective.

In addition, to the extent possible, exemplars should reflect the test-taker population with respect to differences in native language, race/ethnicity, cultural background, etc. This is particularly important in tests with responses that either reveal characteristics related to the test taker (video, spoken responses, etc.), in which demographic subgroup identity can be inferred or in tests with patterns or response types that can be linked to certain subpopulations. The subgroups of interest will differ by testing program as the demographic groups represented in the test-taker population will be specific to the test.

Evaluation and Evidence:

- Document the process for the exemplar selection, including the consideration of the score scale representation, various response types, the process for searching for these response types, and the process for determining consensus scores.
- If appropriate, document how the exemplar selections include multiple and sufficient representations of different test takers' responses/approaches, especially for subgroups that are associated with certain response styles.
- In constructed responses that reveal demographic characteristics of the test taker, document how the demographic composition of the test takers whose responses were used as exemplars represents the demographic composition of the test-taker population, if possible.
- Evaluate exemplars for fairness issues, including bias, sensitivity, and accessibility.

4.2.2 Incorporate Exemplars Representing Atypical Responses

Provide exemplars for raters that demonstrate how to appropriately score responses deemed difficult to score.

Raters will encounter responses that are unusual, whether they are off-topic or simply unscorable. Raters must be trained on how to score them or when to defer them for scoring by a more expert rater.

Evaluation and Evidence:

- Document the process and actions taken to find, review, and select responses that are atypical.
- Perform quality assurance checks to ensure that atypical responses are handled appropriately.

4.2.3 Exemplar Consensus Scores

Ensure that the consensus scores assigned to exemplars are based on a true consensus of content experts and reflect the gold standard application of the scoring rubric and guidelines.

Using an exemplar without a clear and definitive consensus score may lead to rater errors in operational scoring. Consensus scores must reflect the judgment of multiple content experts, including test developers and/or senior scoring staff/leaders who are expert raters with experience and who have consistently met performance standards.

Evaluation and Evidence:

- Document the names and qualifications of those on the panel used to establish a consensus score.

4.3 Development of Training and Scoring Materials for Raters

Rater training and operational scoring materials should be clear, precise, and aligned with the tasks to be scored.

Materials such as scoring rubrics, instructions, annotations, and other materials used for training or supporting operational scoring should be aligned to the tasks for which responses are collected. Aspects of the materials considered incorrect, confusing, extraneous, and distracting may contribute to construct-irrelevant variance in the scores. Further, training materials should be free of bias. When developing materials, these principles should be considered, and a quality control procedure should be in place to ensure they are followed.

One area of emphasis for training materials, regardless of the assessment context, is guidance and training to help raters identify and manage their own sources of personal bias. Raters need materials to support spending time considering their own sources of bias, and to recognize and minimize bias from influencing their evaluations.

Evaluation and Evidence:

- Document the quality control procedure for evaluating all materials used in the training and scoring process.
- Document that the quality control procedure was implemented for new tasks and for the creation of new training and scoring materials.

- Ensure and document that the training and scoring materials include multiple and sufficient representations of different test-takers' responses/approaches.
- Conduct reviews of training materials for tasks that have been shown to be potentially problematic because they yield relatively lower scores or are reported as difficult to score.

4.4 Rater Selection

Recruit raters who meet the minimum hiring standards for the testing program and who represent a wide range of demographic groups.

Raters should be recruited to meet certain minimum requirements such as education level, subject- and/or language-specific knowledge, and previous teaching experience, if applicable. These requirements will be test-specific, as each task may require special skills and abilities. In addition, when appropriate, there should be special efforts in place to recruit raters from traditionally underrepresented subgroups of the test taker population. Diversity in the rater pool can introduce broad perspectives that can mitigate biases.

Evaluation and Evidence:

- Check and document that minimum hiring standards were followed.
- Document the rater recruitment process, including special efforts to recruit underrepresented groups in the current rater pool with the goal of achieving diversity in the rater pool with respect to gender, race, and ethnicity, and other demographic groups, as appropriate.
- Conduct an ongoing evaluation of the diversity of the rater pool. The evaluation should summarize the distributions of major demographic groups.

4.5 Rater Training

Raters should be adequately trained before scoring and provided with appropriate support materials to guide their judgments.

Training allows raters to become knowledgeable about the scoring rubrics and how they are applied in practice, as well as familiar with the operational scoring mode (e.g., online, face-to-face). Time for training should be provided before initial scoring, after long periods of time between scoring sessions and after a rater has been flagged for poor performance levels. Testing programs must define the duration of time in between scoring sessions that would require retraining. Training materials should be available to raters at all times and raters should be encouraged to review and refer to the training materials as needed.

Additionally, raters should undergo bias training designed for raters and be assessed for proper practices after training. This is especially important in testing programs for which raters can infer the demographic characteristics of the test takers.

Evaluation and Evidence:

- Survey raters (new and established) to indicate that the training they received was perceived to be adequate and provided sufficient preparation for scoring.
- Document that training materials are comprehensive and available to raters when needed.
- Document that all raters received bias training.

Case study 2: Because raters were required to have a college degree, but not necessarily in a field relevant to the content area (as noted in Table 2), training materials included prompt-specific content knowledge to help supplement a rater's knowledge of the social studies content. Experience told us that in broad subject areas, such as the social sciences, someone with expertise in one area might not have had significant content knowledge in another (e.g., knowledge of U.S. history but not Economics).

4.6 Raters' Demonstration of Scoring Competence

4.6.1 Rater Qualifying Tests

Prior to operational scoring, use a method to qualify raters that requires them to demonstrate their ability to score reliably at a predetermined level of accuracy.

Raters should be qualified and demonstrate their ability to score accurately, prior to operational scoring, with the goal of ensuring that they consistently produce accurate operational scores. Qualifying tests (e.g., the test a rater must pass to be permitted to score for a specified period of time) should be administered on a regular basis to ensure that raters have maintained their ability to score accurately. Often these qualifying tests are referred to as calibration tests because they serve to reset raters' decision making to align or calibrate raters with the scoring rubric and each other.

The qualifying test may be administered immediately following training and/or at a specified time prior to the start of operational scoring. For example, a program may require raters to qualify for scoring before every scoring session or at some other interval. The required frequency may be dependent upon several criteria, such as the nature of the test, the scoring method used, the stakes associated with the use of the scores, and the experience level of the raters. If a rater is not successful on a first attempt at qualification, it is common to allow a second attempt. If a rater fails qualification, possible actions include dismissing the rater from the scoring sessions or having some form of remediation/retraining before attempting qualification in the future.

Evaluation and Evidence:

- Document the process for qualifying raters that includes the specifications for when and how often the test is administered and the criteria to be used for passing. Document the actions taken if a rater does not pass qualification.
- When possible, collect empirical evidence to support the frequency of the qualifying test administration used by a testing program.
- Analyze and document the outcomes of rater qualifying tests.

4.6.2 Development of Qualifying Tests

Use a method to qualify raters that requires them to demonstrate their ability to score accurately prior to operational scoring.

The qualifying test should be designed to provide evidence that raters understand and internalize the scoring rubrics and can score accurately across all score points. This test may take different forms depending on the nature of the task to be scored. The test may ask raters to score responses to the same task they will score in that current scoring session, or in the case that the scoring rubric is generic and does not require specific content knowledge, the test may consist of responses to a different task that is scored in the same manner as the task they will score in that current scoring session.

If appropriate, the composition of the qualifying test should reflect all score points or at least points that are most prevalent and provide an opportunity for the raters to be assessed on the full range of the kinds of responses they would see in operational scoring sessions. The correct "answers" to the test should be consensus scores.

The reliability of the qualifying test should be evaluated. In addition, the properties of the test should be examined to ensure proper coverage of the score scale and to ensure that one test is not significantly different from other tests, with respect to the difficulty level of scoring. Qualifying tests should also be reevaluated over time to make sure they are still functioning as intended.

Evaluation and Evidence:

- Document that the content of each qualifying test was evaluated by test developers and expert raters (e.g., evaluate "content" specifications, including the representation of the score scale).
- Document that the psychometric properties of each qualifying test were evaluated (e.g., task analyses).
- Document the process used to derive consensus scores for the responses used in the qualifying test.
- Estimate decision accuracy and decision consistency of the qualifying test pass/fail as reliability evidence.

- Estimate correlations between qualifying test scores from different occasions as validity and/or reliability evidence.
- Estimate correlations between qualifying test scores and accuracy rates from the same scoring session as validity and/or reliability evidence. These estimates will need to account for the restriction of range.[5]

4.7 Rater Monitoring

4.7.1 Real-Time Rater Monitoring

Monitor raters in real or near-real time during operational scoring sessions and provide feedback to ensure continued scoring accuracy.

The monitoring of rater performance is critical to ensure that scores are reliable and valid for their intended uses. Ideally, rater monitoring should be conducted as close to real-time as possible (e.g., actively during the scoring session) to give a rater feedback quickly. During a scoring session, senior scoring staff may conduct rater monitoring with of a set of practices that range from very basic to very comprehensive methods. These practices include:

- Using raters' scores on embedded pre-scored sample responses to estimate scoring accuracy, known as validity response monitoring.
- Using expert raters, such as scoring leaders, who score responses and compare their score to those assigned by a rater, known as backrating.

In validity response monitoring, senior scoring staff compare raters' scores on validity responses to the consensus scores given to the validity responses to provide real-time feedback on their scoring. Validity responses are preselected responses that are scored by multiple content-matter experts (test developers and/or senior scoring staff) to obtain a consensus score. Because these responses lead to measures of scoring accuracy that help to provide validity evidence for the functioning of the rater and scoring rubric, they are often referred to as "validity responses" (Wolfe, 2014). This type of monitoring is generally used to assess the accuracy of raters and can be particularly helpful in diagnosing raters' issues using the score scale.

Due to practical issues and logistics, it may be impossible to embed validity responses into operational scoring for raters; in such cases, senior scoring staff rely heavily on backrating. The information about rater behavior gleaned from backrating is commonly captured and acted upon during the scoring session but could occur shortly after the scoring session and be used

[5] Corrections could be made using methods such as those in Thorndike (1949). Alternatively, they could be estimated via a special study in which raters who fail the qualifying test then score under operational conditions. These scores will be combined with scores of raters who passed the qualifying test and used to estimate the correlations between scores on the qualifying tests and operational scoring. This study will remove the restriction of range that would be caused by excluding raters who failed the qualifying test.

to provide feedback at a later time. Furthermore, the senior scoring staff are comparing the raters' scores to their own scores, which are not considered consensus scores.

There are various options that could be implemented for real-time monitoring. The options should be carefully chosen to ensure that the level of monitoring is matched to the stakes involved in the test scores. For example, for low-stakes tests it may not be important to monitor raters as closely as it would be for high-stakes tests. This may influence decisions such as the rate at which validity responses are embedded into operational responses or the number of raters each scoring leader is assigned. These monitoring schemes should be based on current data (e.g., rater scoring rates and rater accuracy information) to optimize the process and ensure that raters are receiving the necessary attention.

To the extent possible, scoring leaders may implement data-driven differentiated monitoring based on metrics such that they target relatively poor-performing raters in a given scoring session.

Evaluation and Evidence:

- At the origination of the testing program and after any changes in the assessment or rater pool, document the process for determining real-time monitoring schemes including the support for the:
  - number of raters on the roster (assigned to scoring leader),
  - validity response insertion rate (the ratio of operational responses to validity responses presented to a rater), and
  - number of papers backrated and the number of feedback messages given to raters by scoring leadership, per scoring session.
- Document the process used for backrating, including instructional materials for raters and suggested backrating practices shared with raters.
- Collect experiment-based evidence showing the impact of different monitoring plans, especially when reductions to the monitoring process are proposed/implemented.
- Collect and analyze backrating and validity response data to evaluate rater effectiveness.

Case Study 1
Human scoring was monitored by scoring leaders during each scoring session using a combination of backrating and validity response agreement. Approximately every 20th response was a validity response which allowed for sufficient data to estimate agreement rates. Scoring leaders were asked to target the amount of backrating, based on validity performance but were asked to backrate at least 1-2 responses per rater per scoring session. Because the rater pool was very experienced and stable, each scoring leader monitored 12 raters.

Case Study 2
Human scoring was conducted during field testing administrations only. Because human scoring was intermittent and we wanted to more closely monitor raters, there was a validity insertion rate of 10% (every 10th response) and the ratio of raters to scoring leaders was 6:1.

4.7.2 Ongoing Regular Monitoring of Individual Raters

Perform ongoing statistical monitoring of individual raters and use compiled performance data to identify raters for remediation, target real-time monitoring efforts, and provide feedback to raters.

Individual rater performance data should be compiled and evaluated over time to better understand a rater's overall performance level. Cumulative performance data should be used to determine if a rater's performance levels do not meet minimum standards. Such raters can then be appropriately remediated or removed from scoring. The timeframe and amount of data collected for this type of monitoring will depend on the context of the assessment, including the stakes of the assessment and periodicity of scoring.

In addition to evaluating the cumulative performance on validity responses, cumulative results based on backrating data recorded from real-time monitoring may be used as a secondary data source to monitor individual rater performance.

When possible, regular feedback should be provided to raters on their cumulative performance even when they are meeting minimum standards. In addition to providing remediation and feedback, cumulative performance data may be used to assist senior scoring staff with real-time monitoring (see 3.7.1).

Evaluation and Evidence:

- Document the process for monitoring individual raters, including the thresholds for classifying raters into performance levels and identifying raters for remediation or dismissal.

4.7.2.1 Rater Remediation and Rating Invalidation

Identify raters exhibiting less than ideal scoring accuracy or behavior and offer opportunities for additional training that targets their deficiencies.

When a rater's values on the evaluation metrics are below a predetermined threshold, they should be remediated/retrained or removed from the rater pool, either permanently (may not score in future scoring sessions) or temporarily (may not score in the current scoring session until they complete additional training and improve their performance on a subsequent evaluation).

Additional actions may be required for scores produced by raters identified as inaccurate, such as checking a sample of the rater's responses for inaccuracies or removing and replacing their ratings, depending on what is feasible and other actions required by a testing program.

<u>Evaluation and Evidence</u>:

- Document the specific conditions for rater remediation and dismissal, as well as the type of remediation that should be implemented.
- Document the requirements for remediated raters or temporarily removed raters to continue rating.
- Document the specific conditions when ratings are to be invalidated, if applicable.

4.7.3 Ongoing Rater Pool Monitoring and Management

Conduct ongoing monitoring of the rater pool using metrics to provide feedback to raters and determine whether corrective action is needed.

In addition to monitoring the performance of individual raters, it is important to monitor the performance of the raters as a group. Evaluating rater pool performance serves as both an evaluation of how accurate/consistent scoring is overall, as well as an important opportunity to identify and diagnose common misunderstandings and issues that the raters may be having. It also serves as a source of evidence for the validity of scores for the test.

For example, if an evaluation of either a validity response scoring performance or an operational scoring session identifies that the raters are, as a group, having difficulty with differentiating between two score points, training sets can be created and deployed to all raters to help correct the issue. Doing this level of analysis helps make remediation as efficient as possible.

Analyses of the rater pool may be performed at different time intervals by different parties. For example, scoring staff may conduct regular analyses and reviews to manage raters' performance. They may rely on some analyses and estimation of statistics that are embedded within the scoring computer systems. Psychometric staff may conduct analyses based on the rater pool on a regular basis after each administration as part of their test reporting procedures and documentation.

If the rater-pool analysis produces results that show unacceptable performance, actions to address those issues should be taken. These actions could include reviewing and modifying rubrics, removing tasks from operational use, supplementing exemplars and training materials for clarification, and reviewing the processes used to train, qualify, remediate, or remove poorly performing raters.

Evaluation and Evidence:

- Document the established rater-pool-evaluation system, which utilizes group-level metrics to diagnose global performance issues and the corresponding actions that would be taken if issues were discovered.

Case Study 1

For this assessment, raters rated every day and its rater pool was established and experienced, with historically stable scoring performance. Consequently, monthly monitoring of overall pool performance was used. In cases where a significant change occurred, such as adding many new raters or the substantial changes to the testing population (e.g., a new client, state, or country adopted the assessment), then monitoring the program temporarily switched to more frequent monitoring to detect and act on issues that had arisen. Monitoring reverted to monthly once the staff were satisfied that scoring quality had stabilized at an acceptable level.

Monthly, rater pool accuracy was calculated using data based on validity responses, and inter-rater agreement, which was calculated using the 5% randomly selected agreement sample of responses that were scored by two raters. Quadratic Weighted Kappa (QWK) was used to assess both agreement accuracy with the validity responses and inter-rater agreement. Details on QWK are discussed in 7.1.2. The data were evaluated relative to historical monthly data of the past two years, to look for potential shifts in the statistics over time, and to see if the data fell into the historical range.

Validity agreement range (QWK): 0.78 - 0.85

Inter-rater agreement range (QWK): 0.73 - 0.83

Data points that fell out of the historical range were flagged for investigation into the scoring process. Data trending in one direction over a period of several months would have also been investigated.

The investigation may have included additional finer-grained analysis, as well as some review and judgment from content scoring experts. It may have uncovered a need for retraining, clarification of a point in the scoring materials, or the need for new samples that illustrated newly emerging test-taker-response approaches.

Annually, QWK for validity agreement and inter-rater agreement were calculated and documented as part of the annual technical report.

Case Study 2

This assessment used human raters only during the field test once per year. The same QWK statistics for validity accuracy and inter-rater agreement as described in Case Study 1 were calculated daily during the scoring period. The thresholds for flagging QWK were 0.8 for validity accuracy and 0.75 for inter-rater agreement.

If flags occurred, additional analyses such as examining score distributions or cross-tabulations of human ratings to evaluate score-point- by score-point agreement, may have been used to identify where the rater pool was having specific difficulty. Additional re-training may have been required by the pool.

If performance was judged to be too weak, score invalidation and rescoring may have been required.

Once field test scoring was completed, QWK for validity agreement and inter-rater agreement was calculated and documented as part of the annual technical report. These statistics were also reviewed by program staff to consider whether changes to procedures were needed or may also have been considered in future item development if the statistics were particularly weak or strong.

4.7.3.1 Estimating Inter-Rater Reliability

Implement a system to track inter-rater reliability on a regular basis.

Inter-rater reliability is a measure of the extent to which raters working independently agree on assigned scores (Lange, 2016). It is sometimes also called inter-rater agreement or rater agreement. There are many ways that the consistency between raters can be calculated (e.g., percent agreement, kappa, intraclass correlation), but all methods rely on a sample of multiple (i.e., at least two) ratings per response. If the scoring design does not require more than one human rating for responses, testing programs should collect scores from multiple human raters for a certain proportion of randomly selected responses in order to estimate inter-rater reliability. This sample is often referred to as the agreement sample. While inter-rater reliability cannot confirm that raters are using the rubric as intended, poor inter-rater reliability is a signal that perhaps the scoring rubrics and guides are insufficient and/or need to be revised to make the scoring process more objective so that raters are using the same process to assign scores.

Evaluation and Evidence:

- Perform and document rater reliability analyses on an ongoing basis, depending on the testing program's administration schedule.
- Document the system for inter-rater reliability estimation including the number of responses to be double-scored, how the second rater is selected within the scoring system, if and how the second rating will contribute to the reported score, how inter-rater reliability will be computed, and what actions will be taken if the reliability estimate does not meet minimum standard**.**

4.7.3.2 Detecting Rater Drift

Determine if raters change their scoring behaviors over time.

It is important to ensure that the rater pool is scoring consistently over time, that is, there is no rater drift. Rater drift refers to the changing (over time) of the implicit standards raters use when making their judgements, which is a threat to the comparability of scores and fairness. There are a few methods for monitoring rater consistency over time. One method requires collecting ratings on previously scored responses (from a past scoring event) that then are embedded in the current scoring queue. The original and new ratings are compared to detect rater drift over time. This process sometimes is referred to as *trend score data collection*. Currently trend scores are collected when tasks are reused on an assessment in a subsequent administration. When tasks are not reused, other methods for detecting drift should be examined such as comparing ratings to other data sources like selected-response-section scores.

Evaluation and Evidence:

- Document the system for collecting trend-score data including the number of exemplar responses to be embedded in operational scoring, how rater drift will be estimated, and what actions will be taken if there is substantial rater drift.
- If possible, perform trend scoring analyses and document the results as well as the actions taken if substantial rater drift was detected.

Case Study 1
Trend scoring studies were conducted annually. A sample of candidates (*n* = 2,000) was selected from the previous three years of test-taker responses to be rescored. The sample was stratified to ensure that all score points were represented (*n* ≥ 25 at every score point) and divided evenly between domestic and international test takers. These older responses were mixed in with the test-taker responses from the current test administration and rescored. The raters did not know which responses were old and which were current.

The raw task-level human ratings were compared using the original and rescored ratings for the older test-taker responses at every point on the score scale (i.e., original ratings of 1 through 5). If the mean of the new ratings was ± 0.25 from the original ratings (e.g., if the mean of the new ratings for responses originally rated 2 was above 2.25 or below 1.75), the item was flagged for rater drift. If rater drift was detected, assessment development staff evaluated and determined what kind of remediation was needed, which might have included additional rater training, clarification of training materials, or some other change.

Case Study 2
Not applicable, human scoring of each item only occurs once, when it is field tested.

4.8 Assigning Responses to Raters

The scoring system should include specifications for the assignment of responses to raters for scoring.

Responses should be randomly assigned to raters. To avoid potential biases, the information provided to raters assigned to each response about the test takers and the testing situation beyond the actual responses should be limited as much as possible. When there are multiple CR tasks on the test, the scoring system should include specifications for how the responses from different tasks are assigned. The contribution of any single rater to a test taker's final scores should be considered when designing the scoring process. For example, if there are many CR tasks on a given assessment, limits might be placed on how many responses a rater can be assigned from a given test taker. This will prevent a test taker's overall test score from being affected by any systematic errors common to the ratings specific to a rater. Moreover, this will also prevent halo effects (i.e., a rater's decision on a test taker's response being affected by that rater's score on other responses from that same test taker). This may yield highly correlated ratings that possibly reduce the ability to distinguish the strengths and weaknesses of a particular test taker (Thorndike, 1920). Online scoring platforms, which are widely used in large-scale assessments, generally are designed to follow such rules when assigning responses to raters.

Evaluation and Evidence:

- Document the design of scoring of multiple CR tasks on a test (if applicable) and the design of rater assignment to responses. The documentation should include an explanation of why the design was selected and how it minimizes rater errors, if applicable. It should also review steps taken to maintain the anonymity of the test takers and their characteristics.

4.9 Maintenance and Incremental Improvement of Human Scoring System and Materials

4.9.1 Ongoing Maintenance of Test and Scoring Materials

On an ongoing basis, evaluate the test and scoring materials currently in use to ensure that they are still aligned with the construct definition and that the consensus scores are still correct.

Exemplars and their consensus scores as well as other rater training materials should be reevaluated on a regular basis and removed when they provide an outdated example that is no longer consistent with the construct definition. For example, consensus scores for exemplars that have been found to be tricky or unsuitable once raters start using them may no longer be valid. Those exemplars should be removed, or the consensus scores be reexamined. In addition, there may be some evolution in the way test takers respond over time, which could also render some exemplars obsolete. Additional training materials should be added to the scoring system on a regular basis.

The CR tasks, should be reevaluated on a regular basis, using analyses of response-ratings data from recent administrations that are representative of the overall test-taker population. For example, tasks with assigned ratings that show very little variability or that yield low rater-accuracy rates should be reviewed and possibly removed from future administrations. Alternatively, poor task performance such as low levels of discrimination may be a signal that exemplars or training materials are not performing as intended.

Evaluation and Evidence:

- Document the process used to reevaluate whether tasks, responses, and consensus scores reflect the construct definition.
- Document that the process is carried out on a regular basis, including findings and actions taken as a result of the evaluation process.

Case Study 1

Statistics for exemplar responses were reviewed quarterly. Validity and calibration responses were evaluated on the basis of percent agreement statistics. Validity and calibration responses with less than 60% exact agreement or 3% or more discrepant ratings were flagged. Since low agreement statistics for an exemplar indicated that its score value was not sufficiently clear, exemplars with low agreement were removed. Again, the values used here were based on experience with similar tests. Flagged responses underwent a qualitative review process similar to what was used for selected response item analysis; a content scoring expert (either assessment development staff or perhaps a scoring leader) reviewed the response in light of the statistics to ensure that the consensus score was correct.

If the score was deemed to be incorrect, the response was removed and replaced. In some cases, exemplars with poor statistics were turned into training or feedback responses with rationales to enhance training if the response was judged too borderline between two score points to be useful for validity or calibration purposes. However, poor statistics alone were not considered a sufficient reason to remove the response from use.

Case Study 2

Exemplar response statistics were reviewed daily during field-test scoring. The same rules and processes described in Case Study 1 were used, but the threshold for 0-1 score scale responses was less than 80% exact agreement and the threshold for 0-2 score scale responses was less than 70% exact agreement or more than 2% discrepant ratings.

Although monitoring occurred daily, actions taken on the basis of statistics for a given response may have been delayed until there was a reasonable sample size (n=50) unless the statistics were judged to be so poor as to warrant examining reviewing them sooner.

4.9.2 Maintenance of Computer Systems

To ensure accurate data collection and storage of ratings, any computer systems used to implement human scoring should be quality tested on a regular basis and when any changes are made to it.

The scoring system distributes responses, captures ratings, and allows raters to evaluate and score test-taker responses. Systems can allow raters to score either virtually or in person. The accuracy of the scoring system is imperative for reliable test scoring. Thus, the scoring system should be updated per best practices for maintaining software, including regular software testing and maintenance. As with a computerized test administration, the scoring system must operate without any software errors or glitches. Technical difficulties must be addressed expeditiously to prevent the assignment of ratings under less than optimal scoring scenarios (e.g., trouble viewing a full essay).

Evaluation and Evidence:

- Document the quality control/assurance guidelines for the scoring system and ensure that they are carried out on a regular basis.
- Document any occurrence of technical difficulties and report on the problem, solution, and impact.

4.10 Final Remarks on Human Scoring

This section (above) on human scoring describes the current best practices which are, in part, based on historical procedures. However, they are also continuously evolving due to both a purposeful scientific reevaluation and client requests/needs. Testing programs may not find all these principles relevant and may adopt other practices that are more suitable for their tests. Of prime importance in the development and maintenance of a human scoring system is that the practices that are implemented should be (to the extent possible) based on empirical evidence or theoretical grounds.

5. Collecting Validity Evidence for the Use of Automated Scores

Automated scoring systems use a variety of computer algorithms, such as natural language processing algorithms, to score written, spoken, or multimodal responses (e.g., videos) for the purposes of providing holistic scores, analytic subscores, diagnostic feedback, routing decisions, or similar kinds of information. An automated scoring "engine" is the machinery within the scoring system that uses these algorithms to carry out a two-step process to assign scores to responses. The first step in the automated scoring process is the extraction of information from the constructed response. In this first step, the information that is extracted consists of elements of the response that might have been established in the relevant field/discipline to be indicative of the construct under measurement. The extraction of values involves a set of computational routines that yield statistical variables of different scale types (e.g., continuous, discrete), referred to as features. For example, depending on the construct being assessed, these features may represent different aspects of writing quality for a writing assessment, checks for specific key words in written responses for assessments of content knowledge, or logical steps in a mathematical solution for a test of mathematics ability. The process of identifying, developing, and evaluating components of the automated scoring system is one foundation upon which automated scoring is conducted. The features that the first step yields become the inputs for the second step in the process.

In the second step, a computer algorithm uses the combination of features from the first step as inputs for assigning a score or scores or other classification(s) to the response. This algorithm is commonly referred to as the automated scoring model. The development of the scoring model is a key element of the automated scoring process. Details of the automated scoring model and of its development process impact the validity of the scores and should be tracked as evidence of that validity.

Minimizing the differences between the predicted values from the scoring model and the human rating is the most common approach for developing the scoring model. However, the selected method for building the scoring model should always be appropriate for the testing program and tasks. Currently, all automated scoring applications used by ETS are built and evaluated on the basis of their ability to predict human ratings. Since other methods are not used at ETS, we have not developed any practices for such methods. Thus, the discussion herein will focus on best practices associated with automated scoring models that are developed to predict human ratings.

5.1 Construct Relevance of an Automated Scoring System

Design and implement an automated scoring system based on the link between the construct to be measured and the system components.

All components of the automated scoring system, including features and algorithms, should reflect the aspects of the construct that are important to be measured. The scoring model will be more sensitive to or give greater weight to the values of some features than others. These weights should be consistent with the relative importance of the aspects of the construct each feature represents. Care needs to be taken to ensure that the differences in automated system scores are determined by differences in construct-relevant measurements and not covariates that are unrelated to the construct. Alignment with both construct-relevant and irrelevant variables should be tested to produce convergent and divergent validity evidence by demonstrating the expected relationships with both types of evidence. Evidence of convergent validity includes positive and strong relationships with construct-relevant variables and evidence of divergent validity includes weak relationships with construct-irrelevant variables. If no clear direct links are possible to make between the features and the construct, then the indirect link to the construct via the human scores used for model development (discussed in the next section) is even more important to establish and document.

Evaluation and Evidence:

- Document the automated scoring model, its components, features, and feature weightings. The weightings might be obtained directly from the model (as, for example, in the case of linear regression) or, in the case of more complex models, through additional analysis of model behavior.
- Document the rationale for how the features and weights reflect or otherwise relate to the intended construct.
- When applicable, have experts in the field conduct a qualitative review of the features used as inputs in the model and how they correspond to the construct definition.
- Conduct analyses that provide evidence of convergent validity such as moderate correlations between the task or CR section score with other related scores.

Case Study 1
A high-level description of the rubric is given in Section 4.1 and the feature set is provided in Table 1. There was a clear connection between the writing task, the scoring rubric, and the features. That is, the features were designed to be aligned with the components of writing in these types of writing tasks, and the specific features that were selected align to the rubric to resemble the human scoring process. Due to the specifics of the task and human scoring instructions, some of the features were weighted much more heavily by the model than others. These included the features *Organization* and *Development* as they related to building an argument in the written response. For this particular task, features involving lexical complexity related to word choice and average word length were less important.

Case Study 2
The scoring rubrics for the short-answer tasks mostly involved checking for content accuracy or providing evidence, so the score awarded corresponds to the level of accuracy or evidence. The scoring engine used a series of features that relied on heuristic descriptions to identify relevant components of text, but not necessarily content matching for accuracy. For some tasks on this assessment, there were features that indicated the presence of specific content-related keywords. The presence of these keywords was posited to relate to the level of accuracy and therefore higher predicted human scores.

5.2 Feature and Automated Scoring Engine Development and Maintenance

5.2.1 Feature Development

Develop computational routines to generate features as intended.

The first step algorithm evaluates specific elements of a constructed response to generate features that serve as inputs for the scoring model. Features may or may not claim to evaluate specific elements of the construct, such as grammar when assessing the writing construct. These features include, for example, low-level components such as syntactic parsers for automatic speech recognition as well as high-level components such as argumentation detectors or grammatical error detectors. These are germane to evaluating a response in an automated scoring system.

The quality of the extracted feature values from the features affects the statistical performance of an automated scoring system. Thus, the components of the algorithm should be developed with computational routines that consistently generate the feature values for the response. Routines should be well documented, should be connected to the theoretical literature on the construct where applicable, and should undergo comprehensive quality control checks. In addition, components should be trained on data that are independent of the data used to build and evaluate the automated scoring model and the data should be representative of the current composition of the test-taker population.

Evaluation and Evidence:

- Document the quality control procedures for checking the component definitions and extraction routines.
- Document the distribution of feature values using analyses appropriate for each type of feature. The evaluation should include a check to ensure that the expected feature distributions are observed.
- Perform statistical analyses that show that the extracted feature values correspond to other measures of the same aspects of the construct, when features are claimed to evaluate specific aspects of the construct.
- Follow current software development best practices to maintain up-to-date and thorough documentation that describes the creation of a new component or changes to an existing component’s definition/routines. Documentation should be written in technical terminology to achieve accurate descriptions useful for other scientists working on component development. Documentation should also include a description in lay terms for other teams in order to facilitate the comprehension of processes performed by the automated scoring system.

Case Study 1
Computational routines were checked using a relevant dataset. Reasonableness of resulting feature values was evaluated using correlations between features as well as qualitative reviews of text responses.

Case Study 2
Computational routines were checked using a relevant dataset. Indicator-based feature values were hand-verified using a series of text responses. Features derived from pattern matching and other linguistic structures were evaluated in a qualitative review as well.

5.2.1.1 Replicability of Feature Values

Computational routines should consistently yield the same values for the same response.

Features generated by each component and corresponding routines should be replicable and produce the same rating for the same constructed response when submitted repeatedly to the same automated scoring system. Failure to show replicability indicates computational errors or other problems that need to be addressed before the component is used in operations.

Evaluation and Evidence:

- Document the tests performed to show that component computational routines yield the same feature values.

5.2.1.2 Human Annotations for Training Components

Human annotations that are used to create and test components should be of high quality.

Some components are tuned for specific applications using human annotations of CRs. This is referred to as training the component. Not all components rely on such training with human annotations, a process in which humans are instructed to document certain linguistic characteristics of a set of texts to train the algorithm.

If a component is trained using human annotations, all key standards for the quality of the human annotations should be met before they are used for training components.

Evaluation and Evidence:

- Document the quality standards for human annotations.

5.3 Automated Scoring Engine Development, Testing, and Maintenance

Base the set of components of an automated scoring engine on empirical research from the appropriate discipline and have as clear a link as possible to key aspects of the scoring rubric.

Automated scoring engines typically include multiple components developed through research in natural language processing and relevant content domains. The components of the automated scoring engine should be evaluated together at various design levels that are relevant to the specific use-context including, but not limited to, tasks, subgroups, administrations, and time windows. The sample of data used for component evaluation needs to reflect the characteristics of the testing population as closely as possible. The components should be reevaluated on a regular basis and might need to be structurally updated whenever the construct coverage of the tasks for which they are used, the performance characteristics of the testing population, or other influential assessment characteristics change notably. Upon updating the automated scoring engine or the addition of new components, a clear description of the automated scoring engine should be provided in a written report, with associated hypotheses on how the new or changed set of components impact the automated scoring model, score distributions, or other outputs.

The sample of responses used to train and test components must be representative of the population of test takers for which the automated scoring engine will be used. Preferable sampling methods include simple random sampling or stratified random sampling, in which the collection of responses is partitioned into groups as defined by subpopulation definitions and a random sample is then drawn from each group. Often historic data are used to build automated scoring engines for use in the future. In such situations, the stability of the test-taker population composition over time should be tested and documented.

Evaluation and Evidence:

- Maintain documentation of the automated scoring engine development, including all processes, empirical findings, thresholds, and interpretations, as well as the results of evaluations of the statistical and psychometric properties of the scoring models, features, and other components of the engine and updates to engine's components. Documentation should include the sample size, dates, and relevant demographic characteristics of the samples that were used, along with the dates and occasions for re-evaluation because of the potential dependency of the components on the sample upon which they were trained.
- Maintain archived datasets used for developing and evaluating the components of the automated scoring engine.
- Maintain documentation of the sample selection process, which includes a clear definition of the test-taker population and evidence that the sample is representative of the population (e.g., the sample is a random or stratified random sample of the population).
- Maintain evidence of the associations among the scores or other classifications from the scoring model and other relevant/criterion variables.

- Produce a report written in lay terms describing any new components added to the set of components for the automated scoring engine, or any substantial changes to any existing components, to explain how those changes align with the construct definition.

5.3.1 Atypical Input Detection

Evaluate the robustness of the automated scoring system to atypical inputs at various design stages that are relevant to the particular use context.

An automated scoring system must be robust to atypical inputs, which are responses that do not follow the expected response format. There are two main sources of atypical inputs. They can be due to atypical characteristics of the response (e.g., canned responses, off-topic responses, responses of atypical length). They can also be due to technical issues which arise either during the test administration (e.g., poor audio quality for a spoken response) or as the response is processed by the automated scoring engine (e.g., a component failed to produce a valid output).

The system’s performance on such atypical inputs needs to be evaluated separately from the general system evaluation. Such evaluations need to be performed on data that are representative of the use-context for which they are designed. Some atypical detection systems rely on statistical models that predict human judgements of whether a response is atypical. When such models are used to detect atypical inputs, standards for criterion variables should be met prior to system evaluation. The thresholds for evaluating the robustness of the system to atypical inputs need to be set based on statistical performance characteristics, historical benchmarks, and policy considerations regarding relative risks.

During system testing, known atypical inputs should be put into the scoring system to evaluate and ensure the accurate identification and routing of atypical inputs for correct handling. Responses identified as atypical should be treated carefully during scoring model development and evaluation, should be examined in light of the score-use context, and may need to be excluded from scoring model building.

Evaluation and Evidence:

- Document the process and rules for detecting and routing atypical inputs.
- Document the process for testing atypical response detection that is conducted during automated scoring engine development and during the development and evaluation of the automated scoring model.

Case Study 1

The scoring engine processed texts and identified those considered atypical inputs before feature extraction. Atypical inputs for this task included responses written in the wrong language, responses that were blank or too brief based on the number of words, and responses that were off-topic. To test the components of the engine that identified these responses, responses fitting these characteristics were selected from the population of texts used for feature and engine development and used to evaluate the engine. This testing was performed on an annual basis or when an upgrade to the engine was scheduled. When an atypical input did not yield the appropriate classification by the engine during this testing, the computational components for atypical input detection were reviewed.

Case Study 2

In this specific case, the short answer text was entered by the test taker within an online system. The main types of atypical inputs were blank responses or responses in the wrong language. To test that the engine identified these inputs as atypical, responses with these characteristics were selected from the population of texts used for feature and engine development and were used to evaluate the engine. If the engine did not classify these responses as atypical, the computational components for atypical input detection were then reviewed.

5.4 Automated Scoring Model Development (Building and Evaluation)

The process of developing automated scoring models almost always includes the building of statistical prediction models to predict human rating from the features extracted from the constructed response. In current practice, scoring models predict human ratings, but there is no reason why other criteria could not be used if the developer deems them appropriate. We will focus on human ratings as the quantity being predicted or the prediction criteria. Prediction models are controlled by a set of numeric values or model parameters. The values for the model parameters are determined so that the predictions of human ratings produced by the model are close to the true human ratings in this sample of data. Multiple specifications for the models and definitions of closeness of the model predictions and the human raters are available for use in practice. The nature of the responses and the experiences of the model development team contribute to the selection of the model specification and definition of closeness. The selected values for the model parameters yield the predictions that are closest to the human ratings in the sample of data given the specification of the model and the chosen definition of close. Selecting the values for the model parameters is known as *model fitting* or *training*.

Model evaluation refers to the process of performing statistical evaluations of the accuracy of the automated scoring model. This is done at different levels of analysis to choose a single or possibly multiple model(s) as candidates for use. The evaluation of an automated scoring prediction model forms a critical piece in the overall design and implementation process of an automated scoring system.

5.4.1 Characteristics of Samples

Use independent samples that are of sufficient size to yield accurate predictions of human scores and are representative of the testing population and intended score-use contexts to build and train automated scoring models.

Depending on the nature of the assessment, automated scoring models are applied across different contexts that can include things such as tasks (i.e., some automated scoring models are applied to only a single task; others are applied across multiple tasks), subgroups of test takers, administrations, time windows or other factors. Samples used in the development of scoring models should be large enough to contain data representative of the relevant contexts. For example, if a model will be applied to multiple tasks, then the sample should include data from multiple items or if it will be used across multiple time windows, then the data should span multiple time windows.

The samples of test takers whose responses are used in training must be representative of the population of test takers for which the automated scoring engine will be used. This means that the sample must be selected in a manner so that test takers with the entire range of test-taker characteristics, including subgroup identity, and the full range of abilities are eligible to be in the sample. The design for the sample must also allow that the full range of responses in the population also have an opportunity to be selected. The sample should also be representative of the types of tasks that will be scored using the automated scoring engine. However, oversampling, in which some groups of test takers make up a greater proportion of the sample than the population, is possible. Subgroups can be defined by their background characteristics (e.g., racial or ethnic group or language group) or the nature of their responses (e.g., responses that receive scores that are rarely assigned in practice, such as the top or bottom score point). When oversampling is used, analyses that use the sample must be designed to accommodate the sampling plan. Oversampling for the training sample has implications for the accuracy of the scoring models for various subgroups of test takers and across the entire population. Oversampling might be used to improve the fairness of scores and equity of the assessment, but that might come at a cost for the accuracy of the scores across the entire population. The implications of these tradeoffs should be carefully considered when the sample is designed. Random or stratified random samples are preferable. When historic data are used to build automated scoring engines for use in the future, statistical evaluations should be performed to ensure the populations from the two time points are comparable. These might include statistical checks on the comparability of distribution of subgroups or human ratings of the responses, or qualitative evaluations of contextual factors such as college admissions requirements or the school curriculum.

The sample of data used for training scoring models should be distinct from the sample used for scoring model evaluation and, if applicable, any samples used as part of the development of system components. In addition, the two samples should be comparable, not only with respect to demographic composition, but also in feature summaries and distributions of responses identified as having unusual properties (e.g., off-topic responses). It may be necessary to oversample smaller subgroups in the evaluation sample to ensure that adequate subgroup sample sizes are achieved to perform subgroup analyses. Analyses would need to account for such oversampling.

Evaluation and Evidence:

- Document the sample sizes, sample demographic characteristics, test dates, and other information to demonstrate that the automated scoring model training and evaluation samples are representative of the test-taker population.
- Document that the samples used in the process are independent, but comparable.
- Archive the datasets used for model building and evaluation in case the automated scoring model needs to be re-evaluated.

5.4.2 Criterion Variable Quality

Evaluate the statistical and substantive quality of the variables used as prediction criteria prior to using them in model building or evaluation.

Prediction criteria are used for scoring model building and evaluating automated scoring models. The criteria can also be used to evaluate total test scores calculated from automated scores alone or from a combination of automated scores and scores from other tasks. Typically, prediction criteria are human ratings, but they could include other measures such as selected-response scores or other relevant test data. Choosing an appropriate criterion variable should be based on its link to the construct of interest, relevance to the target population, and the measurement quality of the criterion. The quality of the prediction criteria defines the upper limit for the quality of the automated scoring model that is built by optimizing the prediction of the criterion.

Since human ratings are typically the basis for both building and evaluating automated scoring models, the quality of the human ratings is critical. Therefore, all relevant practices outlined in Section 4 (Human Scoring) of this document should be followed for producing the human scores used as the criteria for training and evaluating the accuracy of the statistical prediction models. Caution and careful consideration must be taken if criterion variables are shown to be the result of relatively unreliable or inaccurate processes to determine if their use is defensible in developing automated scoring models in a specific assessment context.

Evaluation and Evidence:

- When using human ratings as the criterion variable, all evaluations and evidence outlined in Section 4 (and listed above) should be conducted/collected and documented, prior to their use in scoring model building and model evaluation.

Case Study 1
The criterion was the human score. Human-human agreement based on agreement sample data was estimated using three years of data. The QWK was .79. Correlations with other sections of the test were strong and positive.

Case Study 2

The criterion was the human score. Human-human agreement based on agreement sample data was estimated using field test data. The QWKs for the five prompts were:

Prompt 1:	0.85
Prompt 2:	0.59
Prompt 3:	0.55
Prompt 4:	0.53
Prompt 5:	0.85.

Note that Prompts 1, 2, and 5 used a 0-2 score scale, and Prompts 3 and 4 used a 0- 1 score scale.

Correlations with the selected response section were moderate to strong and positive for all prompts except for Prompt 4 (the correlation was 0.35).

5.4.3 Appropriateness of Modeling Techniques/Algorithms and Selection Criteria

Statistical methods used to develop and select the statistical prediction models should be appropriate for the properties of the data.

Statistical prediction models should be trained with algorithms that are suitable for the properties of the data (e.g., variable types, distributions, and associations).

Several alternative model specifications will often be compared as part of the development of the automated scoring model. Typically, the specification that yields the most accurate predictions (predictions that are closest to the human ratings or other criteria) is then used. When the scoring model building process requires comparing several models, this comparison should be done using a separate training sample or other statistical approaches such as cross-validation on the model building sample. The evaluation sample should only be used to evaluate the final selected model.

There are various statistical methods used for establishing a prediction model – some as simple as multiple regression, others much more complicated including non-linear models, deep neural networks, machine learning, and other approaches. If applicable, statistical algorithms that allow for a statistical and conceptual explanation of the mechanism for final score prediction should be used, unless other more complex modeling approaches yield substantially more accurate predictions.

Evaluation and Evidence

- Document the different modeling methods explored (if applicable), a justification for why the selected technique is appropriate for the data under analysis, and the technical specifications used for estimation.
- Provide documentation to support the use of more advanced/complex modeling strategies.
- Create a report that explains the methodology in lay terms.

Case Study 1
Non-negative least squares regression was used, which reset the regression weights to 0 for features with negative weights from ordinary least squares regression. Because this engine was established and had been used for several years, and there were only 10 features which had all been shown to be linked to the construct and related to human scores, typically no weights were set to zero.

Case Study 2
Because of the large number of features that were available in this engine, machine-learning methods were explored. Several approaches including lasso regression, linear support vector regression, support vector regression, ridge regression, and elastic networks were run and compared. They were also compared to the standard non-negative least squares regression model. Despite it being relatively more difficult to explain in lay language, support vector regression was substantially better than non-negative least squares regression and it was selected as the final prediction model for all prompts because it best minimized the prediction error. Future prompts will use this model going forward because our research has shown it to consistently be the best at minimizing prediction error for the prompts on this test.

5.4.4 Model Evaluation Analyses

Model evaluation analyses should yield sufficient evidence that predicted scores provide an interchangeable proxy for human scores/judgments that lead to fair conclusions about test takers.

The scoring model evaluation process involves several analyses to understand the quality of the scores from the automated scoring model under consideration. If the testing program is developing an automated scoring system from scratch, the performance of the scoring model will be evaluated using direct comparison to standards from the field such as whether the correlation between human ratings and automated scores exceeds a given threshold such as 0.70 (Williamson, Xi, & Breyer, 2012). However, when the model under evaluation is part of a potential upgrade to the scoring model and engine, much of the analysis will involve comparisons of the performance of the new model to the performance of the existing scoring model. In general, scoring model evaluation analyses might include some or all of the following:

- Differences in human and automated engine-score distributions
- The concordance of the human ratings and the automated scores. Various statistics, such as Pearson correlation coefficients and quadratic weighted kappa (QWK; Haberman, 2019), can be used for this purpose.
- The proportional reduction in mean squared error (PRMSE) when using an automated score to predict a human true score (Haberman, 2008).[6] Mean squared error (MSE) is a common measure of the accuracy of a prediction such as an automated test score from a model fit to predict human ratings. Large values indicate large discrepancies between the prediction and the criterion being predicted (e.g., the human scores or human true scores). MSE is dependent on the scale of the data. PRMSE removes the dependency on the scale by focusing on the improvement in accuracy or the reduction in MSE. It measures the change in the MSE that results from using the automated scoring model to assign scores compared with assigning the average human rating, which would be the simplest prediction model. The average human rating is not a viable model for automated scoring, but it provides a baseline for interpreting the relative accuracy of scoring models.
- Another possible statistic for evaluation that is related to PRMSE is the correlation between the automated score and the human true score. This correlation cannot be calculated directly since the true scores cannot be observed. However, the correlation between the automated scores and the human ratings equals the correlation between the automated score and the human true score multiplied by the square root of the human inter-rater reliability. Since the human inter-rater reliability is less than one the correlation between the automated scores and the human ratings is attenuated as an estimate of the correlation between the automated score and the human true score. The correlation between the automated score and the human true score therefore can be estimated by dividing the correlation between the automated score and the human ratings by the square root of the human inter-rater reliability. This statistic is often referred to as a disattenuated correlation.[7]
- When the new model is an upgrade to the existing model or automated scoring engine:
    1. Differences in score distributions for the new and current scoring model
    2. A comparison of PRMSE for the new and current scoring model

Details on concordance statistics used in these comparisons and their thresholds are provided in Section 7.

Importantly, analyses should be conducted at the rating level (comparing human and automated scores of individual responses) and section-score level (comparing section-level scores based on the aggregation of task scores computed using human ratings for each of the

[6] The human true score is the expected value of all possible human ratings made by trained raters of a response. It cannot be observed but its variance can be estimated under certain assumptions about the human ratings. For further discussion of the role of the true score in evaluating automated scoring see Loukina et al. (2020).

[7] This is also referred to as deattenuated correlation.

tasks versus scores computed using the automated score for each of the tasks). The section-score level comparisons would require that all the CR tasks in the section were double human scored for a subset of test takers. Further, total scores should also be examined when feasible and appropriate, for example, if the test is a mixed format test. It is important to note that although section-score level and total-score level impacts are important to measure and understand, they are insufficient because individual item differences matter with respect to construct relevance. The opposite is also true; analyses at the item-level are generally insufficient, and thus a complete analysis will consider all levels of measurement. However, the criteria used to evaluate scores for an individual task might depend on the contribution of the task to the total score. Different criteria might be used when the task is one of many than if the task is the only task on an assessment as might be true for an extensive writing task.

In addition, if the model is being built for a generic task type (in which one model is used for all tasks of the same type) it is useful to evaluate the model at the task-level, which can possibly reveal some deeper issues with the model or issues with the task (possibly outdated/disclosed). For example, if a task were accidentally disclosed to the public, thereby impacting the distribution of human scores, the prediction error for that task may change.

Evaluation and Evidence:

- Conduct analyses that demonstrate that the automated scoring model predicts the criterion within a predetermined/acceptable amount of error (thresholds are discussed in Section 7).
- Document the analyses performed that evaluate the automated scoring model and write a summary of all evidence demonstrating whether the automated scoring engine is adequate for operational scoring.

	Case Study 1	**Case Study 2**
Modeling Context	• Engine/model upgrade evaluation compared two automated scoring engines which differed in this case by their feature sets. • Generic Model (for 1 task)	• New set of prompts, therefore new models were built • Prompt-specific models (5 prompts)
Differences in human- and automated engine-score distributions	The machine-score standard deviation (SD) was smaller for both engines (current engine: 0.70; upgraded/proposed engine: 0.75) than the human score SD (0.96). At the score point level (considering rounded machine scores), this difference in SD related to more machine scores in the middle whereas there were more human ratings at 1 and 5.	The machine score SD was smaller than the human score SD for all prompts. With respect to means, the standardized mean differences by prompt were: Prompt 1: -0.148 Prompt 2: 0.004 Prompt 3: 0.112 Prompt 4: -0.109 Prompt 5: 0.013

	Case Study 1	Case Study 2
The concordance of the human ratings and the automated scores.	The correlation between the upgraded/proposed engine with human scores was .78 whereas for the current engine the correlation was .72. In this case, the correlation was exactly equal to the QWK.	The QWKs describing the agreement between the machine and the human scores for the five prompts were: Prompt 1: 0.50 Prompt 2: 0.63 Prompt 3: 0.60 Prompt 4: 0.56 Prompt 5: 0.86
PRMSE/MSE	The PRMSE / MSE for the upgraded engine were 0.76 / 0.36 and for the current engine were 0.66 / 0.44.	The PRMSE / MSEs for the five prompts were: Prompt 1: 0.28 / 0.44 Prompt 2: 0.68 / 0.23 Prompt 3: 0.66 / 0.18 Prompt 4: 0.59 / 0.15 Prompt 5: 0.82 / 0.21
Disattenuated Correlation	The squared disattenuated correlation for the upgraded engine was 0.76 and for the current engine was 0.66.	The squared disattenuated correlations for the five prompts were: Prompt 1: 0.30 Prompt 2: 0.67 Prompt 3: 0.67 Prompt 4: 0.61 Prompt 5: 0.83
The concordance of the scores from two automated scoring engines	The correlation between the scores from the two engines was .992. The difference in engine score distributions was an absolute value of .05% or smaller at each score point.	N/A
Overall Remarks/Summary	The two automated scoring engines were very similar in their prediction of human ratings (based on their QWK statistics). The two sets of machine scores were very strongly correlated. The proposed engine had a larger PRMSE which indicated a superior prediction of the human true score, thus there was support for upgrading the engine to use the new feature set which added one feature.	Prompt 1 had poor results overall including a relatively large standardized mean difference between the machine and human scores, low QWK, and low PRMSE. Prompt 4 also had relatively less than desirable results. The model evaluation results for the other prompts did not yield any major concerns.

5.4.5 Comparison of Prediction Model Performance by Subgroup

To ensure fairness, evaluate the model performance for all relevant subgroups.

To evaluate the level of fairness of the automated scoring model to consistently predict human scores, analyses should be performed for all relevant subgroups. Subgroup analyses may include:

- A comparison of the human rating and automated score distributions by subgroup
- A test for subgroup differences in automated score distributions for responses with the same human ratings
- Human-automated scoring engine agreement by subgroup
- Automated scoring engine-automated scoring engine agreement by subgroup (for automated scoring engine upgrades)
- Automated score distributions by subgroup
- Comparison of feature distributions by subgroup
- Differential score performance by subgroup

Importantly, the selection of subgroups to be examined varies by testing program. There should be careful consideration of the proposed uses of the test scores and test-taker population when selecting relevant subgroups for analysis.

Some subgroups will not be adequately represented in the samples and, therefore, due to small sample sizes, they may have a relatively large prediction error. Thus, the results of a fairness analysis might be unreliable. This limitation will impede the extent to which subgroup analyses will be helpful in assessing and ensuring fairness.

<u>Evaluation and Evidence</u>:

- Document the subgroups selected for analysis, the specific analyses to be conducted, and the minimum required sample sizes to be considered. The documentation should include a discussion of the subgroups considered for analyses and an explanation of why they were considered.
- Conduct a fairness evaluation of the automated scoring model for subgroups of interest. Building and evaluating scoring models separately for different subgroups can be useful as an exercise to see how differently the models fit for groups. This exercise could be helpful to evaluate if and how different groups might be systematically advantaged or disadvantaged by a model that is built by a pooled group of test candidates. Differences between the models could indicate differences in the way different groups respond to items that could have an influence on the scores they receive from an automated scoring model.
- Document the results of the subgroup analyses and fairness evaluation in a summary report.

Case Study 1
Extensive analyses were performed to compare relevant subgroups for which there was sufficient data. These analyses examined human-machine agreement (focusing on QWK, Pearson correlations, standardized mean differences (SMDs), and other metrics. The SMDs for these relevant subgroups ranged from approximately 0 to 0.1.

Case Study 2
Subgroup analyses were performed by race/ethnicity, gender, and socioeconomic status. Generally, there were small differences, except for two prompts in the race/ethnic subgroup analysis. One standardized mean difference (SMD) comparing the human and machine scores was .12 and another one was .145. The SMD of .12 was judged to be acceptable because it was a small group and a review of the item did not reveal any substantive fairness issues. It was decided not to use the other prompt (Prompt 1) to compute reported scores because the wording of the prompt was potentially troublesome as per a qualitative review, the subgroup sample size was large, and the relative impact on the reported scores was substantial for a large portion of the subgroup. This prompt also had poor statistics overall. All other SMDs were less than 0.1.

5.5 Automated Scoring Model Monitoring and Maintenance

Determine operational quality control procedures to monitor the performance of the automated scoring model both prior to and after the system is implemented.

Performance standards must be set for an automated scoring model prior to the use of the model for operational scoring. A model needs to be thoroughly evaluated to ensure its accuracy and fairness prior to its deployment, and the entire scoring process should be evaluated in terms of its ability to properly identify and route atypical responses for special consideration. In addition, the model should be monitored continually to ensure that its performance remains acceptable according to the agreed-upon performance standards set by the testing program. Investigation is needed when a model is not performing acceptably, with the goal of rectifying the issue.

In some cases, monitoring will detect issues that are not directly related to the automated scoring model, but rather issues that have arisen in other parts of the assessment system. A comprehensive testing plan that is executed ahead of an administration or after changes or updates to any part of the system is advised. However, even with extensive testing ahead of time, monitoring during the operational scoring period is also suggested.

Evaluation and Evidence:

- Document the performance standards and monitoring/testing plan. The monitoring plan may include:
  - comparing score distributions to expected or historical data,
  - looking for higher or lower proportions of certain score points, or very high proportions of responses that are not scorable by the automated scoring model, or,
  - evaluating the agreement (or an agreement sample) of human ratings on the operational responses.
- Perform and document spot checks and evaluations of the responses that were not scorable by the automated scoring model to determine if the model requires updating.

5.6 Implementation and Ongoing Maintenance of the Automated Scoring Software and Architecture

The implementation and maintenance of the automated scoring software should follow up-to-date software industry standards and best practices.

An automated scoring system is a complex piece of software, and its implementation should follow industry best practices (e.g., those outlined in Spolsky, 2004) and relevant aspects of current industry standards (e.g., IEEE, 2014, p. 730). The automated scoring software development process needs to ensure that the computer code is as error-free as possible and can be maintained and updated by multiple developers. This includes appropriate procedures and tools for documentation, version control, code review, and issue tracking. Comprehensive testing of the software and its apparatus need to be done at various levels including unit testing, functional testing, regression testing, and performance testing using industry standard tools. Significant changes to the code should trigger appropriate automatic testing to ensure that it did not introduce any unintended negative consequences.

Evaluation and Evidence:

- Document plans for following industry best practices for system and software updating and maintenance.
- Document any changes to the software using appropriate industry practices including version-control numbers.
- Document any computational changes, especially related to potential, unintended negative consequences to test takers.
- Document that all third-party software libraries used in the automated scoring software were properly vetted to ensure that they do not introduce potential security, legal, or other threats.

5.7 Utilize Human and Automated Scores in Ongoing Quality Control Measures

Use human and automated scores to perform quality-control checks and detect scoring trends.

If the scoring system incorporates both automated and human scoring, it is advantageous to use each type of rating as a check on the other. A comparison of human and automated scores can help identify either changes in the human scores over a duration of time or an issue that causes changes to automated scoring output.

One advantage that automated scoring has over human ratings is that an automated scoring model will score identical responses with identical scores over time, unless the scoring model has changed. There is no such guarantee with human raters. Human scores are prone to drifting over time, whether by the pool of raters gaining experience, a subtle change in training, or other factors. These score shifts can cause problems of comparability of scores over time. The consistency of the automated scoring engine can be used to identify these score shifts. In this way, automated scores can assist in the quality control of human scores.

Conversely, in an ongoing system it is important to use human ratings to perform reasonableness checks of the automated scoring system. Human ratings will be helpful in detecting errors or issues with the engine that previously did not occur or were undetected (e.g., new types of responses that cannot be scored or responses unlike any in the data used in building the automated scoring model). In this way, human scores can assist in the quality control of automated scoring.

Testing programs should consider such comparisons and plan to monitor score quality. In addition, plans should be in place to make adjustments for human score drifts, such as through score equating.

Evaluation and Evidence:

- Document a comprehensive scoring quality monitoring plan that encompasses both human and automated scoring.
- Document appropriate score equating/scaling/linking plans or some other process that accounts for human score drift over time, particularly if tasks are reused.

5.8 Generalizability of Automated Scoring Models

Account for potential generalizations of automated scoring models beyond the evaluation conditions and sample.

Automated scoring models can be applied in different scenarios such as by task, for specific administrations, or by time windows. The scenarios in which the models will be applied must be clearly stated before the evaluation is conducted. The evaluation must provide evidence that scoring models perform adequately across all levels for which they are to be applied. As a general principle, care must be taken to not overgeneralize results from the analyses (e.g., generalizing from a few tasks to all tasks, generalizing from data from unmotivated field-test takers to operational data).

Evidence and Evaluation:

- Document how the automated scoring model aligns with the structure and assumptions of the testing program (target test-taker population, ratings, tasks, subscores, task-specific or generic rubrics, etc.). The documentation should also explicitly discuss scenarios for which the automated scoring model should not be applied.

6. Evidence for Reported Scores

In Sections 2, 3 and 4, best practices were described for planning, designing, executing, and documenting a scoring process that contributes to a validity argument that uses ratings from either humans or automated scoring. The current section outlines the principles and decisions that need to be considered when creating scores from the ratings. When designing constructed-response tasks and the overall assessments, it is important to consider how the ratings that are assigned to the constructed-response tasks will be used to create scores and their intended uses.

The evidence discussed in this section will build upon all the previously collected validity evidence to support the reporting and use of final scores (whether total or subscores) for their intended purposes. However, it is not necessarily the case that tests will be developed linearly starting with the development of tasks to the decisions on how to report final scores. For instance, an iterative process of decision-making between all the various stages might be necessary, as an assessment is developed and the characteristics of different pieces of the assessment are known, or as circumstances or needs of an assessment program change over time.

6.1 Scoring Mode and Design

The scoring design is a set of rules that governs the rating of constructed responses and all related decisions. These rules include the mode for rating the response (human, automated, or a combination of both); the number of human ratings per response, provided any will be used; and the roles for both human and automated scores when both are used. The determination of the design should consider multiple factors and the process used to make the determination should be documented. The following sections provide additional details on the scoring mode and design.

6.1.1 Determine the Mode for Rating Constructed Responses

The choice of mode for rating constructed responses — human, automated, or both — should be informed by several considerations.

When designing a new assessment and developing its scoring design, one consideration is the mode of scoring: human rating, automated scoring, or both. The selected mode can affect the reliability of the resulting scores for the assessment and the validity of inferences made from those scores. Hence, there are several factors that should be considered when selecting the modality, and support for the chosen mode should be documented. Factors that should be considered include the availability of automated scoring systems for the type of construct and tasks, the evidence in support of the validity of scores provided by those systems, the accuracy of human ratings, and the nature of the assessment (e.g., whether the assessment is used to make consequential decisions for the test taker). The decision to use one type of scoring or a combination should be carefully considered and the reasoning behind the decision should be well documented. Cost may also contribute to the decision, and how it influences the decision should be documented as well.

Evaluation and Evidence:

- Document the types of ratings that will be used, with support based on the description of the proposed use-context, including the role of the constructed-response task(s) in the assessment and the use of the scores, and the evidence gathered about human and/or automated-scoring-system performance.

6.1.2 Scoring with Human Ratings Only

Carefully consider how many ratings are needed when scoring with humans only.

When all of the ratings of a CR task or tasks will be conducted by human raters, a key design issue is the number of human ratings. Almost always, additional ratings of each response will increase the reliability of the final score for the task and the assessment (Brennan, 2001; Haertel, 2006). Hence, the number of human ratings used for each response can be chosen to achieve the level of reliability the assessment requires to meet its claims.

Data with at least two human ratings for each response in a sample of responses is necessary to assess the reliability of scores. Even if the expected plan is to use a single rating, a representative sample of responses will need to be rated by two or more raters each. The raters used for this sample should also be representative of the rater pool in terms of characteristics likely to affect their ratings such as experience and training. The data can then be used to assess the reliability of scores for the task and the assessment under different numbers of ratings for each response. Different methods are possible for assessing reliability under alternative scoring designs using psychometric techniques such as generalizability theory or through stratified alpha coefficients.

If the scores based on human ratings are not sufficiently reliable, even with the available number of ratings, then an alternative scoring design or other changes to the tasks may be necessary.

Evaluation and Evidence:

- Document the process for determining the targeted reliability of the final scores.
- Document the results of an empirical evaluation that compares scoring designs that involve different numbers of human ratings.

6.1.3 Scoring with both Human Rating and Automated Scores

Determine how human and automated scores will be used in combination.

In order to compensate for the shortcomings of either mode of scoring, an assessment design might call for the use of both human ratings and automated scores. Automated scores can be used to improve the reliability of scores compared with using only human ratings. Alternatively, human ratings may better match the construct than automated scores. There are two common ways for using the two scoring methods together. One approach uses the automated scores as a “check” on the human ratings. If the difference between an automated score and a human rating for a response exceeds a threshold, then an additional human rating is collected. The second approach uses the automated scores as contributory scores by taking the weighted average of the human and automated scores to produce a score for a response (Breyer, Rupp, & Bridgeman, 2017).

The choice among these approaches will depend on the evidence to support the use of the automated scores and the reliability of the human ratings. The combination of both human ratings and automated scores will tend to be more reliable than using automated scores as a check. Hence, when the evidence supports the use of automated scores then contributory scoring is likely to be preferred. When the evidence is less clear, then check scores might be useful since they can potentially improve the accuracy of human ratings by catching some poor ratings. If the evidence is too weak, the use of human scores alone might be appropriate.

Evaluation and Evidence:

- Document the plan for how scores will be used. Include all ratings, by type (human and automated), and how they contribute to the scoring process. If there are multiple tasks on an assessment, the rules and process may differ between tasks. Include a rationale for decisions if multiple, reasonably defensible options exist.

6.1.4 Scoring with Automated Scores Only

Ensure that the use of scores based solely on automated scoring can be supported by a validity argument.

Scores generated only by automated scoring should be reported operationally only if validity evidence supports this use and if the risks for potential misinterpretation of the scores as equivalent to human ratings is low. A challenge to using only automated scores is the detection of atypical responses. Human raters can detect certain types of atypical responses. The standard scoring models used to produce automated scores might not be able to detect these responses. Additional algorithms are needed to check for such problems. Hence, if the scoring design calls for only automated scores, then systems for detecting atypical responses will also need to be developed and tested and the evidence of their effectiveness should be included in the documentation supporting the scoring design.

Human rating for a sample of responses is advised for the ongoing evaluation of the automated scores.

Evaluation and Evidence:

- Document the procedures used to evaluate the quality of the final scores to support their validity and the results of these evaluations. The evidence should be comparable to what was considered acceptable for the testing program under the typical human-score-based scoring system (or under a system combining human and automated scores).

- Document the plan for collecting human ratings to provide validity checks for the automated scores. The plan should include the schedule for collecting the ratings and running validity checks, as well as the number of responses that will be scored.
- Document the quality of the atypical response detector algorithms.

6.2 Ancillary Ratings

Determine the role of ancillary ratings in score calculation.

Ancillary ratings are those ratings that are assigned to a response for non-operational purposes. For example, a percentage of responses may be randomly selected to be assigned an additional human rating for reliability estimation, or a scoring leader backrates a rater and the backrating is captured and recorded. These ratings serve a primary purpose for monitoring the scoring process and may be referred to as ancillary ratings. Whether or not to use ancillary ratings in score calculation is an important question. In one scenario, the randomly selected responses receive additional scrutiny that can result in more reliable scores than the scores of responses that do not receive the additional rating. In an opposite scenario, ancillary ratings are not used and thus not all available information is used to calculate the task score for those randomly selected responses. The decision of whether or not to use ancillary ratings depends on beliefs about the importance of equivalent scoring designs for all test takers and the benefits and costs of using procedures that, by design, yield different reliability for the scores of different test takers.

If ancillary ratings are used in score computation, the context in which they were collected should be carefully considered. For example, if the statistics used for the psychometric properties of the assessment scores (e.g., inter-rater reliability) assume rating independence, then it may be important to ensure that the rating is *blind*, that is, the scoring leader did not know the original rating.

Evaluation and Evidence

- Document the costs and benefits/trade-offs of using ancillary ratings versus not using them.
- Document all scoring rules involving ancillary ratings, providing a rationale for each decision that is made.

Case Study 1

Agreement sample ratings were not used for score computation.

If a response was selected to become an exemplar, subsequent ratings assigned during calibration or validity scoring were not used for score computation. These ratings were considered part of the quality monitoring process, not the operational scoring process. Further, dozens or even hundreds of ratings could have been assigned to a single exemplar response long after final scores for the test taker were already reported.

Case Study 2

No ancillary scores were collected during operational scoring.

6.3 Ratings Resolution and Scoring Rules

Determine how to use ratings and calculate scores.

When more than one rating is applied to a response, rules for resolving any differences in the ratings and calculating the final task score must be established.

When there is sizable disagreement among the ratings of the same response whether from multiple human raters or a combination of human raters and automated scores, and they are discrepant by more than some predetermined value, an additional rating may be collected in an adjudication process in order to complete scoring. Otherwise, if those two ratings are in sufficient agreement, then ratings collection is complete.

Once the ratings have been collected, *scoring rules* determine which ratings will be used and how they will be combined to calculate task-level scores. Some of these rules may be very straightforward, such as taking the sum or weighted average of all of the ratings that were collected. Other rules are more complicated. Some scoring rules will discard all other ratings when a rating is available from a senior rater such as a scoring leader or use a subset of the ratings in an average or sum. In some cases, excluding a discrepant rating after the adjudication process is not advised (Cohen, 2015; Mazzeo, Schmitt, & Cook, 1987). All reasonable variations of the adjudication processes and scoring rules should be carefully considered. For more guidance on setting discrepancy thresholds for adjudication and score combination rules, see Breyer et al. (2017).

Whenever scores from human ratings and automated scoring models are combined for a given reporting purpose, different weighting schemes should be evaluated for their statistical performance and conceptual defensibility. The risks associated with using certain combinations of human and automated scores must be identified.

The psychometric properties of the scores computed using different approaches should be analyzed to determine the extent to which each set of adjudication and scoring rules yields scores that can be supported in a validity argument for the task/test. As with ancillary ratings, the importance of equal treatment of the test takers might also factor into the selection of a scoring rule. Furthermore, when changes to ratings' resolution rules are proposed, analyses to evaluate the potential impact on reported scores are advised.

Evaluation and Evidence:

- Document the adjudication process and scoring rules, including the evidence supporting these rules and the associated risks.
- If applicable, produce documentation describing the selected weights that will be applied to the multiple ratings for a constructed response. Also provide a theoretical explanation and empirical analyses that support how the selected weights align with the psychometric needs of the assessment (e.g., maximizing score reliability).

Case Study 1: If the human and automated scores differed by more than one score point, the response was sent for a second human rating. The average of the three ratings was used for calculating scores.

Analysis of discrepant ratings in this program over time suggested that averaging all three ratings produced the most accurate and reliable task score. This was also supported by the literature on discrepant human scores (Cohen, 2015).

Case Study 2: No discrepancy resolution was used.

6.4 Score Transparency

Ensure transparency in the procedures used for producing reported scores.

Test takers, score users, and experts in the field should be informed about how reported scores are created. Descriptions should be understandable and meaningful, especially those procedures that use automated scoring systems, which can involve complex computer algorithms and that lack the intuitive transparency of human rating (i.e., a human reviewing a CR and evaluating it according to its rubric). This includes understanding the strengths and limitations of using human raters, automated systems, or a combination of the two as part of operational scoring. This level of transparency is consistent with the Council of Europe's *Recommendation on the Human Rights Impacts of Algorithmic Systems* (Council of Europe, 2020).

Evaluation and Evidence:

- Prepare and publish documentation that includes high-level descriptions of the human scoring process and the automated scoring systems. For example, rubrics may be published as well as descriptions of rater training processes. For automated scoring, a high-level description might include a list of the features and the types of statistical or artificial intelligence (AI) models that are used and the methods used to train them. The documentation should be readily understandable to the lay public.

- Publish documentation in the form of a technical report or other format with greater technical details.

Case Studies 1 and 2
Technical reports were published that described the human scoring system and the full scoring process, which included the automated score model building and evaluation.

7. Evaluation Metrics and Thresholds for Decision-Making

A prominent source of evidence for constructed-response task scores to support the claims and desired inferences of an assessment is the degree to which scores meet a set of statistical criteria. We use the term, evaluation metrics, to refer to statistics and other data summaries or information collected and used specifically for the purpose of evaluation and to support decisions about the acceptability of the quality or performance of human ratings or automated scores. Typically, these metrics will also include thresholds or other specific criteria to be used in the evaluations and decisions. Some examples of statistics used as evaluation metrics include correlation coefficients, QWKs, and mean squared error. The strength of the evidence typically depends on whether the values of these statistics exceed a target or thresholds. The statistics and thresholds are used to support decisions that are needed during the operational scoring process (e.g., whether to use human raters versus automated scoring to score responses or whether the discrepancy between two sets of ratings requires attention). Such decisions are made on an ongoing basis every time the assessment is administered, and the resulting responses are scored. Other decisions will be about the entire scoring process (e.g., whether to use automated scoring operationally or whether every response should be rated by one or two human raters). Such decisions might be made only one time when the system is updated, or its continued use is evaluated.

This section describes principles and best practices for the selection of evaluation metrics and the derivation and use of thresholds for decision-making. Key to these choices are factors related to the statistical properties of various potential metrics such as the match between the statistic and the inferences needed about the ratings or the precision of the statistic given likely available sample sizes. However, other factors should also be taken into consideration. These factors might include the stakes of the test for test takers and score users and the consequences for the testing program of decisions based on the metrics and thresholds. For instance, if the decision is whether to use automated scoring operationally, then the feasibility of the assessment program to meet test takers' or test users' needs without the use of automated scoring would need to be considered. If needs are more adequately met without automated scoring (for example, with human scoring), then this would need to be considered in any decision.

In general, decisions should be made based on a collection of information rather than a single factor, like whether a metric exceeds a threshold. Using multiple data points to make decisions is consistent with traditional statistical and measurement decision-making practices (e.g., Standard 12.10, American Educational Research Association, the American Psychological Association, & the National Council on Measurement in Education, 2014, p. 198) and the ITC Guidelines on Test Use (Guideline 2.8.4, International Test Commission, 2013). The selection of metrics and thresholds should therefore also be informed by the full body of evidence that is planned for the decision-making process.

7.1 Metrics for Evaluation

7.1.1 Selection of Metrics

Select appropriate metrics for the purpose of the evaluation.

Metrics typically will consist of statistics that describe the psychometric or statistical properties of constructed-response scores or ratings. These metrics should be selected and evaluated in relation to general scientific principles, historical benchmarks, and policy implications in consideration of the types of risks to the validity of intended inferences and score uses of the CR scores.

Human scoring and automated scoring models should be evaluated using statistical metrics that are suitable for the plans for using the scores in generating the final reported scores for the assessment. The metrics for each step of the assessment process should be selected carefully to align with the intended pieces of validity evidence that they represent. They should also be readily interpretable and be updated or replaced whenever the scientific state-of-the-art for these metrics' changes. Importantly, multiple metrics should be used for decision-making, when possible.

Evaluation and Evidence:

- Document that selected metrics are suitable for the types of questions that need to be answered and appropriate for the data that are collected.
- Create a plan to review metrics on a regular basis to determine if there are any changes in the scientific state-of-the-art for these metrics and update or replace them if more appropriate methods/metrics are found.
- Document metrics in publicly facing documents. Documentation should describe each metric and its relevance to the decisions it is meant to inform. The documentation should be readily understandable to a broad audience.

7.1.2 Use of Metrics in Human-Rater Monitoring

Use metrics that are appropriate for the type of data as well as the quantity of available cases.

As described in Section 4, the evaluation of human ratings involves monitoring both individual rater performances and the performance of the rater pool. Hence, metrics are needed to support both types of evaluations. Metrics used to evaluate and diagnose *individual raters* include measures of the raters' accuracy when rating validity responses. These metrics should be based on a sufficiently large number of cases to ensure that the precision of estimated statistics is adequate for their intended uses. This may translate into enforcing a minimum number of validity response scores per rater to compute the concordance between the rater's scores and the validity scores. The frequency with which raters are monitored may be dependent on the testing program and the number and frequency of administrations during the year.

Commonly used metrics for monitoring the *rater pool* typically involve the consistency of raters when rating the same responses. The metrics include agreement rates and measures of inter-rater reliability such as kappa, QWK, or the intra-class correlation (Haberman, 2019). Exact and adjacent agreement rates are widely used as metrics in the evaluation of human ratings because of their simplicity and intuitive appeal. However, agreement rates are highly sensitive to the number of score points and the distribution of scores so that simple guidelines used to specify "high agreement" do not actually imply equal accuracy of the raters for all tests. Other metrics that account for chance agreement such as kappa or QWK or even the cross-tabulation of ratings are likely to be preferable in many settings.

Much like when using evaluation metrics for individual raters, care should be taken to ensure that there are sufficient data from which to estimate relevant metrics and make decisions about the overall performance of the rater pool. In addition, the timeframe or duration of time from which to collect and use rating data to estimate metrics should be specified to ensure that the evaluation metrics are current and not based on outdated information.

Evaluation and Evidence:

- Document the process by which a testing program decides on the statistics to be used to evaluate individual raters and the rater pool, and the guidelines for using those statistics in decision-making (minimum sample size, timeframe, etc.).

7.1.3 Selection and Use of Metrics in Automated Scoring Model Evaluation

Metrics should be chosen in relationship to the validity argument to be developed for automated scores and the assessment of the statistical properties of scores.

For the evaluation of automated scoring models, considerations for selecting appropriate metrics include:

- What comparisons between the selected criterion variable (e.g., human ratings) and the automated scores will be valuable for demonstrating the validity of the proposed scores for their intended purposes? What methods and statistics will support those comparisons and desired inferences?
- What comparisons between automated scoring engines (e.g., between an existing and modified engine) or automated scoring models (e.g., two potential models of which one will be used for operational scoring) are required for making determinations about the possible use of the proposed scoring model? What methods and metrics will support those comparisons and determinations?
- What postprocessing will be applied to automated scoring system-generated scores, including rounding, in the reporting of scores? How can the metrics be matched to that processing? For example, if unrounded scores are used in operational scoring, system scores should not be rounded during the evaluation.

In automated scoring, the scoring model evaluation must use suitable statistics to determine the accuracy of the scores produced by the scoring model. Commonly used statistics include the correlation, QWK, MSE, PRMSE, and disattenuated correlation. The appropriate method for calculating some statistics depends on the data and the score scale; consequently, for some statistics, multiple formulae exist, and care must be taken that the appropriate formula is used. Some formulae are appropriate for only specific, limited applications and do not generalize across other applications. For instance, the most commonly reported formula for QWK assumes discrete data but alternative formulae exist for continuous data such as some automated scores (Haberman, 2019).

Some automated scoring models produce continuous values even though human raters produce discrete ratings. In these cases, a decision must be made whether to convert the automated scores to discrete values (e.g., by rounding) for calculation of statistics used in the evaluation. In general, continuous values will be used to create evidence of validity. When considering reliability and impact on the assessment scores, the automated scores should be discrete if that is how they will be used in the calculating the reported score(s). That is, if rounded values are used when calculating the assessment score, then rounded values should be used in the evaluation of reliability and impact on scores. If continuous scores are used, they should be used in the evaluation. Similarly, sometimes automated scores are rescaled, for example to have the same mean and standard deviation as human ratings. Again, the scale used in the calculation of the assessment score should be used in the evaluation.

The selection of metrics should consider the required sample size to ensure sufficient statistical precision to support the decisions based on the metrics. If the necessary samples cannot be obtained, then alternative metrics should be considered. Methods for calculating standard errors and confidence intervals should also be considered when metrics are chosen.

Evaluation and Evidence:

- Document how the selected metrics support the validity argument. Documentation should be detailed and include computational formulae and guidelines for appropriate interpretation.
- Documentation should also include evidence for the precision of the statistics and consideration of the minimum sample size for the calculation of the statistics.

Case Studies 1 and 2

Several metrics were used in model evaluation to help understand how well the engine predicted human scores and how similar they were with respect to their distributions. In both case studies, there were sufficient data to estimate a series of statistics including QWK, PRMSE, MSE, and SMD. Large PRMSE and QWK values and small MSE and SMD values supported the use of the machine scores as they indicated that the machine scores sufficiently reflected human scores which, in turn, supported the validity argument.

7.2 Thresholds for Evaluation Metrics

Thresholds are used throughout the test/task development and scoring process to classify statistics used as evaluation metrics into levels for decision-making purposes. With respect to human scoring, for example, thresholds may be used to monitor and identify raters who fail to use the entire score scale, or who tend to have bias towards high or low scores. In the automated scoring context, thresholds may be used to classify the performance of scoring models as fitting well (acceptable) with minimal prediction error or fitting poorly (unacceptable) with excessive prediction error, and the levels in between. Following from the above definition, thresholds will inherently be specific to the statistic that is used. Further, since statistical criteria/metrics may differ by testing program (depending on the type of score scale and assumptions), any two testing programs or even different tasks with a testing program may use a different set of metrics and thresholds (Williamson, Xi, & Breyer, 2012). There is no "one-size-fits-all" when it comes to thresholds. Further, the purpose of setting thresholds is not intended to establish gates for making "go/no-go" decisions, but rather to create checkpoints to help make more comprehensive decisions in a larger context of validity, fairness, and reliability of scores.

The actual thresholds adopted and the way in which they are used will vary due to differences in the following:

- *Properties of the statistics* – Some statistics used for evaluation can be sensitive to the distribution of scores or the number of score points for the item. The thresholds might need to be adjusted for the distribution or score points. For example, agreement rates tend to be higher when the scores are concentrated and there are few score points and therefore may not be a good choice for an evaluation metric. However, if they are used, the threshold should reflect the score distribution and score points. QWK tends to be lower for a short score scale (e.g., 0-1 and 0-2) than for scales with more points (Brenner & Kliebsch, 1996). For items with short scales, thresholds for QWK might be set somewhat lower than for items with more possible score points.
- *Stakes of the test* – Thresholds for high-stakes tests will likely be more stringent than thresholds for low-stakes tests or formative assessments.
- *Use context* – Thresholds will vary for the same statistic when it is being used in different scenarios. The thresholds will tend to be more stringent in context with greater reliance on the constructed-response scores. In automated scoring, these may include automated scoring model building from scratch (when a testing program has only used human scoring), an automated scoring model or automated scoring engine upgrade, or a sample refresh (when the same set of features are used in the automated scoring engine but the scoring prediction model is trained on a new sample of test takers). For human scoring, these may include qualifying for entering a rater pool, qualifying for a scoring session, ongoing performance evaluation, or evaluation post remediation.
- *Relative "cost" of different decisions* – Thresholds will vary based on the cost of the decision. For example, thresholds for rater termination and rater hiring will not be

symmetric or equally extreme because the risk of firing a good rater is lower than the risk of hiring a potentially bad rater.

- *Evidence for decision-making* – The stringency of the thresholds for metrics used in a decision-making process might depend on the expansiveness of the collection of evidence available for decision-making. Less stringent thresholds might be appropriate when there is more evidence that can be considered holistically.
- *Client influence/contract/restrictions* – The thresholds may be set by the client.

Evaluation and Evidence:

- Document the process to establish/set thresholds for all metrics in the human and automated scoring processes for a testing program. This documentation should be compiled at the origination of the testing program or at the point at which a metric and/or threshold is introduced/revised.
- Produce empirical study results that show the selected threshold results in the intended outcomes. If the metric is being used to make direct assessments about individuals, evidence should demonstrate that the threshold results in a fair and valid decision.

7.2.1 Setting Thresholds for Human Scoring

Thresholds for human rater evaluations should be determined under consideration of the metric, historical performance of the threshold for the metric, and empirical evidence.

The human scoring process involves metrics and thresholds for a series of decisions such as qualifying for scoring, rater remediation, and the overall evaluation of the rater pool. These decision points start at the origination/design of a scoring system and continue to the final evaluation of score quality. Some examples of thresholds in the human scoring process include:

- Passing score for qualifying tests
- Minimum inter-rater agreement statistic
- Minimum rater-level validity agreement
- Minimum item-level validity agreement
- Minimum overall validity agreement for the test
- Maximum amount of drift in scores

Agreement statistics for human scoring may be simple exact agreement rates, exact plus adjacent agreement rates, and discrepancy rates. Other measures of concordance such as QWK may be used. To establish thresholds for decision-making it is best to consider: the metric that will be used for scoring evaluation, the historical thresholds used for that metric, the design of the assessment materials, and the results from empirical studies.

The first step is to consider the statistic that will be used for the metric and understand the properties of that statistic, including the probability distribution and empirical sampling distribution relevant to the testing program. It is important to distinguish between metrics capturing similar properties, such as different agreement rates, and consider how differences in those metrics and how they are computed might require different thresholds. For example, when establishing thresholds for a measure such as QWK, whether it is acceptable to use the same threshold for kappa, weighted kappa, and QWK should be based on how those metrics differ computationally/statistically and in their assumptions. In addition, properties of the score scale should influence the setting of thresholds. For example, QWKs for shorter scales are expected to be lower than QWKs for longer score scales (see Brenner & Kliebsch, 1996). The opposite is true for percentages of exact agreement between raters – shorter scales are expected to have higher percentages of exact agreement.

In addition, historical thresholds may be a primary consideration, especially if there is a client restriction or a psychometrically related restriction that will not permit a deviation from historical practices. Importantly, historic thresholds should be researched to understand the outcomes they have yielded and project the outcomes they might yield, given the future use in a different testing program or context.

One note of caution in the use of historical thresholds is that it is very important to consider the characteristics of the tasks, particularly if newly redesigned tasks are compared to old tasks, or if thresholds from other assessments are used. For example, thresholds may be appropriate under one set of conditions but may be too stringent or too lenient in a new context. The consideration of differences between contexts must guide the use of historical precedents.

To the extent possible, the best practice for establishing thresholds should be empirical studies evaluating the implications. For example, an empirical study might ask the question: How different would rater accuracy be in the rater pool if the passing threshold on the qualifying test to hire raters was 80% versus 70%?

7.2.2 Setting Thresholds for Automated Scoring

The goal of the empirical evaluation of automated scores is to provide evidence to support the validity argument for the final test score and specific claims that the scores are a meaningful numeric evaluation of the skills the item intends to elicit and assess. The empirical evidence should include evidence of the fairness of the item.[8]

The evaluations will differ depending on the design of the items. Some items are designed and developed so that the human scoring rubric specifies the link between the performance and the assessment of skills or content related to the construct. Furthermore, the rubric specifies how to provide a numeric evaluation of the performance. Alternatively, items could be designed with the link between the performance and the assessment of the skills specified in terms of the automated scores with a specification of how the scoring model will produce a numeric evaluation of the performance.

[8] As noted previously, the discussion of automated scoring assumes that tasks were designed with the expectation that human raters would provide scores. Tasks designed with the expectation of automated scoring without reference to human ratings will be considered in the future as the use and practice around such tasks increase.

Human rater reliability: There is no specific requirement that the reliability of the human ratings be above some threshold in order to use the ratings to develop and evaluate the automated scores. Very low reliability (e.g., inter-rater reliability of around 0.30 or less) calls into question the ability of the raters to use the rubric properly and this might suggest the need to improve human ratings. Low human inter-rater reliability will degrade the prediction accuracy of the fitted prediction model for a fixed sample size. For instance, in applications at ETS, we have found that when the QWK for human-rater agreement is below 0.60, the fitted prediction models are nearly always too inaccurate to support the use of the automated scores. Lower reliability necessitates the use of a larger sample to achieve the same accuracy as higher reliability. Low human-rater reliability will also degrade measures of concordance between the automated scores and human raters such as QWK or correlation. For this reason, measures of the concordance between the human true score and the automated scores are preferable. These measures include PRMSE or the disattenuated correlation between the human ratings and the automated scores.

Concordance with the human rater true score: The correspondence between the automated scores and human ratings is a key element of the numeric evaluation of automated scores for items with human-rater rubrics. The evaluation requires that best practices are followed for the development of the items and the training and supervision of the human raters. There should be evidence that the human raters rate the constructed responses in accordance with the rubric, that human ratings (as a group) are not biased, and that human raters are assigned responses to rate in a manner that is effectively random. Under these assumptions the concordance with the human-rater true score provides validity evidence for the use of automated scores.

Subgroup differences: Any evaluation of automated scores must also include checks for differences in the performance of the scores across different identifiable subgroups of test takers that could threaten the fairness of the scores. A common practice is to the compare mean of the automated score to the mean human rating for each subgroup. Heuristically, the automated scores should not advantage or disadvantage a subgroup on average relative to the human ratings. The difference in means is typically standardized by dividing it by the standard deviation of the human ratings for the entire sample or square root of the average of the variance for the human ratings for the entire sample and the variance of the automated scores for the entire sample.[9] Following Williamson et al. (2012), thresholds for the subgroup differences commonly are set to small values such as 0.10 or 0.15. Often subgroup sample sizes are small. Confidence intervals for the differences should be also be considered. An alternative concept of fairness for automated scores is that equivalent responses (i.e., responses reflecting equivalent ability levels) from students of different subgroups receive the same score from the automated scoring engine. Loukina, Madnani, and Zechner (2019) define equivalent responses as those with equal human ratings. However, because human ratings are not perfectly reliable,

[9] At ETS, we also consider the differences in the subgroup means of the standardized z-scores for human ratings and the standardized z-scores automated scores. The z-score for the human rating equals the $(H - \text{Mean}_H)/S_H$, where Mean_H equals the mean for the entire sample of human ratings and S_H equals the standard deviation for the entire sample of human ratings. A similar formula is used to calculate the z-scores for the automated scores.

it may be more appropriate to treat responses with equal human *true* scores as equivalent. Johnson and McCaffrey (2021) propose methods to test for subgroup difference under this approach. These methods for checking fairness are relatively new and so there are not common rules of thumb used for thresholds, but a threshold of 0.10 or possibly 0.15 for appropriately standardized estimates of subgroup values is likely to prove useful. Establishing thresholds for these methods is an important component of future research.

7.2.3 Guiding Questions for Setting Thresholds for Concordance Statistics

Three questions guide the setting of thresholds for the empirical evaluation of the concordance between the human ratings and the automated scores.

1. How important is the concordance with human ratings, as a source of content evidence, such that the automated scores measure the content or construct the item is claimed to measure, according to its specification?

   All else equal, thresholds for the concordance statistics will tend to increase with the importance of the concordance evidence. The following factors contribute to the importance of the concordance with human ratings as a source of content evidence.

   - Item design and content evidence from the design. The concordance is of high importance for items designed for rating by humans and for which the rubric and the assumption of accurate human judgements are central to the content argument for the item. Alternatively, the concordance with human ratings is less critical for items designed explicitly to match the affordances and limitations for automated scoring and for which the qualitative content evidence acknowledges the role automated scoring will play.
   - Other sources of content evidence. The existence of other sources of content evidence reduces the importance of the concordance with human ratings. These sources can include evidence that the input features extracted from the response are related to the content or construct to be measured by the item. This evidence can be qualitative, based on the theory about the construct and the feature design, or empirical by showing that feature values correlate with criteria (e.g., human ratings) in expected ways. The scoring model, that converts feature inputs into scores, can also provide evidence that automated scores are related to the content. In general, models that are less transparent (e.g., complex AI models like deep neural networks) provide less content evidence for the automated scores. Convergent and discriminant evidence from the relationship between the scores and criteria other than human ratings on the same responses also reduces the importance of concordance with human ratings.

     The evidence on the features, model, and the associations with other external criteria can also be used to evaluate the sources of variance in the automated scores and the potential contributions of construct irrelevant variance and the associations of that variance with other important attributes of the test takers such as background characteristics. The evidence can also identify construct underrepresentation or the potential for misleading inferences and unintended consequences resulting from it.

2. How are automated scores used in the creation of the final reported score and how close of an approximation to human ratings does this require?

   The proposed role for the automated scores in determining the final score for the item and the test can affect how closely the automated scores should approximate the human ratings. All else equal, situations that require closer approximations to human ratings tend to require a higher threshold for the concordance statistics. The following factors contribute to the importance of a close approximation to the human ratings.

   - Automated scores can be used as direct substitutes for human ratings in the sense that in items designed and originally rated by human raters, human ratings are replaced by the automated scores without any other changes to scoring of the overall test. Any equating functions, conversion tables, or IRT or similar models used or developed using human ratings will be used with the automated scores without modification. These cases require very high concordance between the human ratings and the automated score. Not only should the human ratings or their true scores and automated scores be highly correlated, the automated scores might need to have a similar mean and variance as the human ratings or have the same distribution across possible score points.

   - Automated scores might also be combined through an average or weighted average with the human ratings. Typically, higher thresholds are beneficial in such cases. Sometimes the weights are defined via statistical models, such as the Best Linear Predictor (BLP, Haberman, 2020; Haberman & Qian, 2007), that adjusts for the correlation between the human ratings and the automated scores and the scale of both. In these cases, the threshold might be somewhat lower than if the weights were set using some other methods such as just using the average or using human judgement to set them. However, other times the weights are set a priori. In these cases, the higher thresholds might still be appropriate.

   - Automated scores can also be used as a check on the human ratings so that if the human ratings and automated scores disagree to too large a value, then a corrective action such as additional human rating is taken. Although the automated scores make no direct contribution to the test takers' final scores, the automated scores must have strong concordance with human ratings. Otherwise using them as checks will be inefficient: problematic human ratings will be missed and human ratings that are truly problematic will not receive corrective actions, Such a case will often not result in improved scores, but will incur costs.

- At the other extreme, human ratings might only contribute to building the automated scoring models. For example, the automated scores may be the only scores used to score constructed response items that would then be combined with scores from selected response items for use in total scores and only the automated scores and selected response scores would be used for equating. In these cases, there would be less need for automated scores to be highly aligned to the human ratings, provided there is sufficient content evidence to support the use of the automated scores.

3. How much damage can be caused by construct irrelevant variance or construct underrepresentation introduced by automated scores?

   The potential for any construct irrelevant variance or construct underrepresentation introduced by automated scores to harm test takers can vary across different potential scoring designs and tests. Two factors affect the potential for harm.

   - The stakes of the test taker and test users. Some assessments are used for more consequential decisions about the test takers than others. For example, school admissions test scores can affect the test takers' educational future and outcomes and they can also have consequences for the institutions since they affect the makeup of their student population. Generally, testing situations with higher stakes require stronger evidence in support of the scores. Hence, all else equal, the thresholds for the concordance statistics will increase with the stakes of the test.
   - The contribution of an automated score from a single item can vary considerably across different assessment. For example, an automated score might be the sole score (no human scores) for the only item or one of very few items on a test (or subsection of a test for which scores are reported). This would be true for a writing test with two items, each scored only by automated scoring. Alternatively, the automated scoring might be combined with human ratings, which reduces its contribution. Likewise, an automatically scored item might be one of many items contributing to the final scores so that the item's score has very limited impact on the total score. Again, all else being equal, the thresholds for the concordance statistics will increase with size of the contribution of a score to the total score.

The answers to these three questions can help determine whether higher or lower thresholds are appropriate, but they do not specify high or low numeric values. The following rules give some general rules of thumb for setting the numeric values.

7.2.4 PRMSE is Very High (PRMSE > 0.95).

If the PRMSE is very high (PRMSE > 0.95), then the automated scores will have statistical properties that are very similar to human ratings for most measurement uses:

a. They will have a mean and standard deviation that are very similar to human ratings;
b. They will work similarly in equating or scaling models;
c. They will have similar correlations with other criteria;
d. They will yield similar IRT parameters;
e. There will be little opportunity for construct irrelevant variance to be in the automated scores, construct under-representation, or bias across subgroups.

The automated scores can be used for the most part interchangeably with human raters. Regardless of the answers to the three guiding questions, scores with a PRMSE of over 0.95 will almost always satisfy the needs of the test. Conversely, if the test design calls for automated scores to be used in situations that require a high concordance, such as serving as a direct substitute or interchangeably for a human rating, then a very high PRMSE is required and 0.95 would serve as a reasonable threshold.[10]

7.2.5 PRMSE is Below 0.70.

PRMSE below 0.70 provides very limited evidence to support claims that the automated scores accurately measure the skills captured in the response, in accordance with the item design and the human-rating rubric. The PRMSE is a function of the square of the correlation between the automated scores and the human true scores, and the means and standard deviations of the automated scores and the human true scores (Casabianca et al., 2021). If the scale and location (i.e., the standard deviation and mean, respectively) of the automated scores are not of special importance, then the square of the disattenuated correlation between the automated scores and the human ratings, which estimates the square of the correlation between the automated scores and the human true scores, might be a preferable metric to consider. Scale and location are generally not of special importance when:

a. The scores for the item will be equated or rescaled;
b. The scores will be used in an IRT model to produce the final score;
c. The scores are combined with human ratings or across multiple items using BLP.

[10] In cases where a very high PRMSE is required, the automated scores should be rescaled and centered to minimize the MSE. That is, the mean of the automated scores should equal the mean of the human ratings and the standard deviation of the automated score should equal the standard deviation of the human true score multiplied by the squared correlation between the automated scores and the human true score. The correlation between the automated scores and the squared human true score multiplied by the variance of the human true score equals the squared correlation between the automated scores and the human ratings multiplied by the variance of the human ratings.

Scale and location are of specific importance if the raw automated scores will be reported to users on the scale score specified by the human rubrics or combined with the human raters with prespecified weights such as via a simple average of human ratings and automated scores. They are also important if automated scores are used as a check against the human ratings.

The value of 0.70 can serve as the minimum threshold for using automated scores in practice, however, the value of 0.70 is not immutable. The support for this value is mostly two-fold. First, the value of 0.70 is often used as a target or minimum threshold for human inter-rater reliability. The inter-rater reliability equals the square of the correlation between the human ratings and the human true scores. Hence, if the square of the disattenuated correlation between automated scores and human ratings (which estimates the correlation between automated scores and the human true scores) is greater than 0.70 then the automated scores and human ratings will be held to the same standard.

Human ratings and automated scores differ in significant ways. Typically, we expect deviations between the true score and the human ratings to be uncorrelated with other criteria or variables. This is not necessarily true for automated scores. Meeting this condition is somewhat of a best-case scenario for automated scores. Consequently, if under best case conditions the automated scores fail to meet the minimum threshold for human ratings, then it is reasonable to say the automated scores also fail to meet minimum requirements.

The second motivation for the use of 0.70 is that it is the value suggested by Williamson, Xi, and Breyer (2012) for both the human inter-rater reliability (or human-human correlation) and the QWK for the human ratings and the automated scores. They also recommend a small degradation between the human inter-rater reliability and the QWK for the human ratings and the automated scores (less than 0.10), and small standardized differences between the means of the human ratings and the automated scores (less than 0.10 or 0.15) are commonly used. At a minimum this combination of requirements would tend to flag similar items as the use of PRMSE of 0.70. Moreover, these or similar criteria have been used productively in practice by ETS and other organizations. Hence, there is some evidence from practice to suggest this threshold prevented serious problems from using automated scores.

It is possible that scores with PRMSE or a squared disattenuated correlation below 0.70 could be used in operations. The special cases would tend to involve cases where the human raters are not used in scoring and not central to the content evidence and for which the item’s scores have limited contribution to the total test score and the stakes of the assessment are generally low. Other evidence might also support the use of automated scores in this extreme situation. Such evidence might include the results of item analysis such as the score distribution not being highly concentrated at any one value or at the extremes and strong correlation between the scores on the item and the total score of the other test items.

7.2.6 Setting Threshold Between the High and Low Extremes.

When setting the threshold between the two extremes just discussed, then a holistic evaluation of the answers to the three questions and the full body of evidence to support the use of the item is needed. This other evidence might include item analysis or positive experience with using similar automated scoring engines and models in similar testing contexts. When the answers to all the questions indicate the need for a high threshold, then values for PRMSE (or squared disattenuated correlation if scaling is not important) near 0.95 are advisable. Likewise, if all answers suggest low thresholds are acceptable then a value of 0.70 for PRMSE or the disattenuated correlation is appropriate. When the answers to the question are less consistent, there is no clear ordering or precedence among the three guiding questions. In any specific application it may be important to decide if any questions should have precedent over the others. Given the holistic nature of the decision and the large uncertain number of factors that might be considered (including those that go beyond the empirical analysis), expert judgement is almost certain to play a role in setting the threshold for any particular application. It is advisable that the answers to the questions be carefully documented and the justification for the selected thresholds be documented. If multiple parties have an interest in the use of automated scores (e.g., contractor and client) then all parties should agree to the thresholds prior to the evaluation to minimize the possibility of evaluation threshold values being selected in part on the basis of the outcome of evaluation analyses.

Case Study 1

1- *How important is the concordance with human ratings as a source of content evidence that the automated scores measure the content or construct the item is claimed to measure according to its specification?*

In this case study, the concordance was not the only source of content evidence. Other evidence included:

- Features were developed with the intent to assess well-defined components of written work that had a theoretical connection to the quality of writing;
- Extensive research was conducted on the features prior to use;
- Evidence from continued use showed relationships to writing and other abilities that were consistent with hypothesized relationships;
- Automated scores were calculated using simple scoring and transparent scoring models and the relative weights the model gave to different features was consistent with expert judgement of the importance of the different components for writing.

In terms of providing content evidence, a lower threshold might have been acceptable.

2- *How are automated scores used in the creation of the final reported score and how close of an approximation to human ratings does this require?*

The automated scores were substitutes for the human ratings and the scores for the writing test were equated or linked using the combination of automated scores and human ratings. However, the score for the item equaled the simple unweighted average of the automated score and a human rating which meant the scores should have been scaled accordingly. PRMSE was the appropriate statistic for the evaluation but how the scores were used in this case study did not suggest high thresholds were necessary.

3- *How much damage can be caused by construct irrelevant variance or construct underrepresentation introduced by automated scores?*

Construct irrelevant variance or construct underrepresentation introduced by automated scores had the potential for considerable damage.

- There were only two tasks on the writing assessment; both were CR tasks and scored using the average of an automated score and a human rating;
- The test was used for consequential decisions about the test takers.

The risk was somewhat mitigated because the automated scores were averaged with the human ratings. It was also mitigated by the extensive experience the testing program had with using a very similar model with this test for several years without any evidence of problems.

Taking all the factors together, the pre-existing research on the features and their theoretical connection to the construct of interest provided content validity evidence to support claims for the automated scores. Evaluation of the transparent scoring model was added to the support by showing that the features received weights that also aligned to theoretical expectations about the relative importance of the subconstruct assessed by the features to the overall construct. Similarly, the data from past uses also provided evidence that the scores supported inference about writing ability. This extensive body of evidence beyond the direct comparison to the human rating, reduced the reliance on the accuracy of the predictions to demonstrate the validity of the scores and would have supported setting the threshold for PRMSE on the lower end of the interval between 0.70 and 0.95. The use of the average of a human rating and an automated score also would have supported a lower threshold. However, the writing score depended on scores from only two items and the assessment was high stakes for test takers. Higher thresholds were more appropriate for these factors. Given that the scores would have been averaged with the human ratings using preset equal weights, the mean and scale of the automated scores mattered, hence PRMSE was the preferred statistic for the evaluation. A threshold of 0.75 could have been used for PRMSE given these factors.

Case Study 2

1- *How important is the concordance with human ratings as source of content evidence that the automated scores measure the content or construct the item is claimed to measure according to its specification?*

In this case study, the concordance with human ratings was a very important source of content evidence since most of the feature inputs were linguistic (e.g., *n*-grams) with little explicit connection to the construct, and the features were task specific so there was no historic experience with the features. In addition, the prediction model was a complex AI model for which the contributions of individual features to the output were difficult to assess and therefore could provide no additional evidence of the content relevance of the scores.

2- *How are automated scores used in the creation of the final reported score and how close of an approximation to human ratings does this require?*

The automated scores were not substitutes for the human ratings. They were used as the sole scores for the items. The total scores for the test and the equating and linking of the total scores used only the automated scores. Therefore, there did not need to be a close approximation to human ratings.

3- *How much damage can be caused by construct irrelevant variance or construct underrepresentation introduced by automated scores?*

The fact that the automated score were the sole score for each task increased the risk from using the score since this increased the importance of the automated score. Also, there was no historical evidence on the risk since the testing program had limited experience with this type of scoring engine or with the performance of the scores from a similar task. However, the risk of harm from using the automated scores was greatly mitigated because each automatically scored task was one of many items administered to each student, CR items contributed only about 40% to the total score, and the test had limited consequences for students and educators.

Taking these factors together, the importance of human ratings for providing evidence of content relevance of the scores suggested using higher values of PRMSE or a squared disattenuated correlation. Similarly, the limited experience with these types of scores and the complexity of the model increased the risk of using the scores and, consequently, the value of using a higher threshold for PRMSE or a squared disattenuated correlation. However, the scores are not substitutes for human ratings and the risk of harm from this item was limited because this item had limited contribution to the total score. Additionally, item analyses would have been conducted on the items to provide external evidence of the scores. These factors supported the use of lower thresholds. Because the automated scores would not have been used as substitutes for human ratings in any way and all the scaling would have been based on the automated scores, the squared disattenuated correlation was preferred. Given all of these factors, a threshold of 0.80 was selected for the squared disattenuated correlation. Tasks with a squared disattenuated correlation greater than or equal to 0.80 were included in the task pool for the test. Other tasks were set aside for further analysis. If the test included fewer tasks or put substantially more weight on the CR task scores and the scores had high stakes, then a squared disattenuated correlation threshold of 0.90 or even 0.95 might have been justified.

7.3 Using Thresholds for Decision-Making

The way in which thresholds for metrics are applied might vary and, thus, the decision-making process should be clear and documented.

In many scenarios, decisions are not made on one datapoint alone. In fact, to the extent possible, multiple statistics should be used in decision-making. In these situations, each statistic typically will have its own established threshold. Individual statistics and their thresholds might not be applied as a strict decision-making mechanism. Instead, thresholds would be used as indicators of possible concern, which must then be weighed against the accumulation of information across all the statistics. For instance, several statistics not meeting their thresholds might signal a clear decision, whereas the appropriate decision would be less clear with a collection of statistics in which some but not all achieve their thresholds. The risks indicated by each metric would then need to be considered when interpreting the collection of results. Importantly, the intended use of the metric and corresponding thresholds should be well understood and documented.

Evaluation and Evidence:

- Document the decision-making process, including how thresholds will be used, either in a hard application (strictly based on the level of the statistic in relation to the threshold) or in a soft application (as an indicator of a possible problem within a context of several statistics, considered together for any decisions).
- Produce empirical study results to provide evidence that the decision-making process results in decisions supporting the valid, fair, and reliable assessment of individuals.
- When multiple statistics are used for evaluation, document the rationale for final decisions about models, task use, and other decision points that were made.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® V. 2. *The Journal of Technology, Learning and Assessment*, *4*(3).

Baldwin, D., Fowles, M., & Livingston, S. (2005). *Guidelines for constructed-response and other performance assessments*. Princeton, NJ: Educational Testing Service.

Bejar, I. I. (2012). Rater cognition: Implications for validity. *Educational Measurement: Issues and Practice*, 31(3), 2-9.

Bennett, R. E., & Zhang, M. (2016). Validity and automated scoring. *Technology and testing: Improving educational and psychological measurement*, (F. Drasgow, Ed.),142-173. Routledge.

Brennan, R. L. (2001). *Generalizability theory.* New York: Springer.

Brenner, H., & Kliebsch, U. (1996). Dependence of weighted kappa coefficients on the number of categories. *Epidemiology*, *7*(2), 199-202.

Breyer, F. J., Rupp, A. A., & Bridgeman, B. (2017). Implementing a contributory scoring approach for the GRE® Analytical Writing section: A comprehensive empirical investigation. ETS Research Report Series, 2017(1), 1-28.

Casabianca, J. M., McCaffrey, D. F., Johnson, M., Ricker, K., Rotou, O., & Martineau, J. (2021). Exploration of the Proportional Reduction in Mean Squared Error for Evaluating Automated Scores. ETS Research Memorandum, in preparation.

Cohen, Y. (2015, April). The "Third Rater Fallacy" in Essay Rating: An Empirical Test. *Paper presented at the National Council for Measurement in Education. Chicago, IL*.

Council of Chief State School Officers & Association of Test Publishers (2013). *Operational best practices for statewide large-scale assessment programs*. Council of Chief State School Officers and the Association of Test Publishers.

Council of Europe. (2020). *The human rights impacts of algorithmic systems.* https://search.coe.int/cm/pages/result_details.aspx?objectid=09000016809e1154

Educational Testing Service. (2014). *ETS standards for quality and fairness*. Educational Testing Service.

Educational Testing Service. (2015). *ETS guidelines for fair tests and communications*. Educational Testing Service.

Educational Testing Service. (2021). *ETS guidelines for developing fair tests and communications.* Educational Testing Service.

Educational Testing Service. (n.d.) *Fairness guidelines*. https://www.ets.org/about/fairness/guidelines/

Fahim, M., & Bijani, H. (2011). The effects of rater training on raters' severity and bias in second language writing assessment. *International Journal of Language Testing*, *1*(1), 1-16.

Haberman, S. J. (2008). When can subscores have value? *Journal of Educational and Behavioral Statistics, 33*, 204–229.

Haberman, S. J. (2019). Measures of agreement versus measures of prediction accuracy. *ETS Research Report Series*, *2019*(1), 1-23.

Haberman, S. J. (2020). Application of Best Linear Prediction and Penalized Best Linear Prediction to ETS Tests. *ETS Research Report Series*, *2020*(1), 1-25.

Haberman, S. J., & Qian, J. (2007). Linear prediction of a true score from a direct estimate and several derived estimates. *Journal of Educational and Behavioral Statistics*, *32*(1), 6– 23.

Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 65 - 110). Praeger.

IEEE. (2014). *IEEE 730-2014—IEEE standards for software quality assurance processes*.

International Organization for Standardization. (2017). *ISO standards catalogue.* Retrieved from https://www.iso.org/standards-catalogue/browse-by-ics.html

International Test Commission. (2 013). *ITC guidelines on test use* (ITC-G-TU-20131008). Retrieved from https://www.intestcom.org/files/guideline_test_use.pdf

Johnson, M. & McCaffrey, D. F. (2021). *Evaluating fairness of automated scoring in educational measurement*. [Manuscript in preparation]. ETS.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological bulletin*, *112*(3), 527.

Lange, R. T. (2016) Inter-rater Reliability. In J. Kreutzer, J. DeLuca, & B. Caplan (Eds.), *Encyclopedia of Clinical Neuropsychology*. Springer. https://doi.org/10.1007/978-3-319-56782-2_1203-2

Leacock, C., & Chodorow, M. (2003). C-rater: Automated scoring of short-answer questions. *Computers and the Humanities, 37*(4), 389-405.

Livingston, S. A. (2009). Constructed-response test questions: Why we use them; How we score them. *R&D Connections, 11*. Educational Testing Service.

Loukina, A., Madnani, N., Cahill, A., Yao, L., Johnson, M. S., Riordan, B., & McCaffrey, D. F. (2020). Using PRMSE to evaluate automated scoring systems in the presence of label noise. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 18–29). Seattle, WA, USA: Association for Computational Linguistics.

Loukina, A., Madnani, N., & Zechner, K. (2019, August). The many dimensions of algorithmic fairness in educational applications. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 1-10).

Mazzeo, J., Schmitt, A., & Cook, L. (1987). The comparability of adjudicated and non-adjudicated scores on the ATP English Composition Test with Essay. *Unpublished manuscript*.

McClellan, C. A. (2010). Constructed-response scoring—Doing it right. *R&D Connections*, *13*, 1-7.

Mislevy, R. J. (1994). Evidence and inference in educational assessment. *Psychometrika*, *59*(4), 439-483.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). A brief introduction to evidence-centered design. *ETS Research Report Series*, *2003*(1), i-29.

Spolsky, J. (2004). *Joel on software: And on diverse and occasionally related matters that will prove of interest to software developers, designers, and managers, and to those who, whether by good fortune or ill luck, work with them in some capacity* (3rd ed.). Apress.

Thorndike, E. L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology, 4,* 25–29.

Thorndike, R. L. (1949). *Personnel selection: test and measurement techniques*. New York: Wiley.

Weigle, S. C. (1994). Effects of training on raters of ESL compositions. *Language Testing*, 11(2), 197-223.

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice*, *31*(1), 2–13. https://doi.org/10.1111/j.1745-3992.2011.00223.x

Wind, S. A., Wolfe, E. W., Engelhard Jr, G., Foltz, P., & Rosenstein, M. (2018). The influence of rater effects in training sets on the psychometric quality of automated scoring for writing assessments. *International Journal of Testing, 18(*1), 27-49.

Wolfe, E. W. (2014). Methods for monitoring rating quality: Current practices and suggested changes. *Iowa City IA Pearson (2014). Retrieved April 17*, 2017.

Glossary

Adjacent agreement
Adjacent agreement typically refers to agreement within the score levels directly above and below the exact score match (if applicable), but also may include additional score levels depending on the number of score points on the scale.

Adjudication
The process in which an experienced human rater evaluates a response that has been assigned scores beyond a maximum allowable difference between multiple humans and/or an automated scoring engine, the results of which are often used to compute the final reported score.

Agreement rate
The percentage of cases for which raters agree with each other, with a consensus score, or with an automated score.

Analytic scores
Result of a scoring method in which CR tasks are scored on separate, predefined dimensions (or aspects) of those tasks.

Ancillary ratings
Ratings that are assigned to a response for non-operational purposes.

Atypical input
An atypical response or invalid outputs as the response is processed by the automated scoring system (e.g., the syntactic analyzer failed to produce a valid output).

Atypical responses
Responses that are unusual, whether they are off-topic or simply unscorable. There are two main sources of atypical responses: atypical characteristics of the response (e.g., unmotivated test takers, responses that cause security concerns, responses of atypical length, responses that try to game the system) and technical issues, which occur either during a test administration (e.g., poor audio quality for a spoken response).

Automated scoring model
A computer algorithm that takes information extracted from the constructed response as input and assigns it a numeric value or score.

Automated scoring system
Software that employs a variety of natural language processing algorithms to score written, spoken, or multimodal responses for the purpose of providing holistic scores, analytic subscores, diagnostic feedback, routing decisions, or similar kinds of information.

Automated speech recognition
A process of analyzing spoken responses that are processed with computational routines that provide scores on various characteristics of the responses.

Backrating
A process during which a more experienced rater (e.g., a scoring leader) evaluates task responses in order to determine whether the scores assigned by other human raters or an automated scoring engine are appropriate.

Best linear predictor
A composite score estimated using both human ratings and automated scores with weights selected to minimize the mean squared error between a criterion, such as the human rater true scores, and the composite.

Bias
In general usage, unfairness. In technical usage, the tendency of an estimation procedure to produce estimates that deviate in a systematic way from the correct value (Educational Testing Service, 2014, p. 54). See ***Fairness***.

Component
An algorithm within an automated scoring system that identifies and/or evaluates specific elements of a response to a CR task on the path to generating features.

Consensus score
The score assigned to a written or spoken response that is based on the collective judgment of a group of expert raters.

Construct/Construct definition
The underlying knowledge, skills, and abilities that are assessed by a test. The construct definition spells out what comprises the construct.

Constructed response
The answer produced by a test taker to a test item, problem, question, prompt, or assignment that cannot be answered by selecting responses from a pre-specified set of choices.

Constructed-response task
A test item, problem, question, prompt, or assignment that requires the test taker to create an answer instead of choosing one from a list of options. The response may be required in one of several possible modalities (e.g., writing, speaking, multimodal output).

Construct-relevant/irrelevant
Factors that influence test scores because they are related/not related to the underlying construct to be measured.

Continuous data
Information that can be of any value within a certain range or scale (e.g., length, height, weight, automated score of 3.4).

Contributory scores
The situation when automated scores are combined with human ratings for the purposes of score reporting to save the costs of additional human raters (Breyer et al., 2017).

Correlation
Correlation measures the strength of association between two or more variables.

Criterion variable
The quantity that is being predicted in a statistical analysis (e.g., predicting human rater scores).

Decision accuracy
The proportion of test takers who are classified correctly by an assessment, based on the data from that assessment.

Decision consistency
The extent to which a decision based on an assessment would remain the same across potential alternative test forms and raters for the assessment.

Disattenuated correlation
The correlation between the human rating and the automated scores corrected for the less than perfect reliability of the human rating. It equals the correlation between the human ratings and the automated scores divided by the square root of the correlation between two human ratings.

Discrepant rating/Discrepancy
A rating assigned to a response that differs from a consensus score by more than a predetermined maximum allowable difference. Alternatively, two scores on an operational response are discrepant from each other when they differ by a predetermined maximum allowable difference. The maximum allowable differences between raters' scores depend on such factors as the total number of score points or score classifications and the use context of the scores.

Discrete data
Information that can be classified into distinct categories or counts is considered discrete (e.g., male/female, yes/no, correct/incorrect, 450 doctors).

Discrimination
A statistic that predicts the probability that a person of a certain ability or higher will answer a specific item correctly and a person of lower ability will answer that same item incorrectly.

Empirical Sampling Distribution
The distribution of a statistic over repeated random samples from a population.

Exemplar responses
CR responses that experts agree illustrates performance at a particular score level; used for human rater training and as part of operational scoring processes. Sometimes referred to as *benchmarks.*

External criterion scores
Scores that come from other assessments or sources that are used for comparative reasons.

Fairness
Fairness is the extent to which the inferences made on the basis of test scores are valid for different groups of test takers (Educational Testing Service, 2014, p. 19).

Feature
Unique classifications describing aspects of a response based on the decomposition of text or speech sounds using components or computational algorithms that can be represented computationally. These classifications can be characteristics such as parts of speech, word counts, syntax classifications, etc. and collectively are used to derive an automated score when they are used as the inputs for scoring models. Depending on the type of automated scoring engine, a feature may or may not have direct ties to a construct.

Feature scores
A calculated numerical representation of the occurrence of features in a text. The higher the feature score, the more important the feature is towards the computation of the automated score.

Field test
The administration of a test still in the development phase used to collect information about the psychometric properties of the test including test- and task-level reliability, validity, and fairness evidence.

Holistic scores
Result of a scoring method that uses rubrics to guide raters in making a single qualitative evaluation of a response as a whole, thereby integrating different aspects of the rubric into a single score.

Human annotation
A process in which humans are instructed to document certain linguistic characteristics of a set of texts which, in turn, are used to train a Natural Language Processing algorithm.

Human rating/score
The value/score assigned by a trained human rater to a test taker’s response.

Internal structure
The internal structure of a test is "the degree to which the relationships among test items and test components conform to the construct on which the proposed test score interpretations are based." (AERA, APA, NCME, 2014, p. 16).

Inter-rater reliability
A measure of the extent to which raters working independently agree on assigned scores. It is also sometimes called inter-rater agreement or rater agreement.

Mean squared error (MSE)
A statistical measure of the average variation in prediction errors for certain statistical models such as multiple linear regression models.

Multimodal response
A response that includes data of multiple formats such as text, audio, or video.

Natural language processing (NLP)
A scientific discipline concerned with the automatic processing of linguistic data (text or speech) in order to make inferences about linguistic phenomena or develop routines for scoring, prompting, routing, feedback, and other relevant tasks in assessment.

Operational scoring
The act of scoring CRs in which the ratings assigned will be used in score reporting to test takers or test users. This term is typically used in contrast to scoring for the purposes of rater training, qualification, monitoring, or the development of training materials.

Pilot testing
Studies in which new items or test forms are administered to small samples of test takers conducted prior to larger scale field studies to get preliminary data to see how well items or tasks perform.

Prediction criteria
The numerical quantities to be predicted by prediction model.

Proportional reduction in mean squared error (PRMSE)
A statistical measure used in the evaluation of automated scoring models to determine the accuracy of predicting human true scores from automated scores.

Quadratic weighted kappa (QWK)
A statistical measure of classification agreement that is adjusted for chance-agreement and penalizes more strongly discrepant classifications using a quadratic weighting scheme.

Qualifying test
A process used to check human rater accuracy before operational scoring; it establishes that raters correctly interpret and properly use the established scoring criteria/rubric. Also sometimes referred to as a calibration test.

Rater accuracy information
Data and summary statistics used to examine how consistently raters' scores agree with the consensus scores on validity responses.

Rater drift
The changing (over time) of the implicit standards raters use when making their judgements, which results in changes in ratings for equivalent responses and is a threat to the comparability of scores and fairness.

Raters
Individuals (sometimes called readers, scorers, markers, or judges) who evaluate responses from constructed-response tasks.

Rater scoring rates
The average amount of time it takes for an individual rater to score a constructed response.

Reliability
The extent to which scores (or other reported results) on a test are consistent across— and can be generalized to — other forms of the test and, in some cases, other occasions of testing and other raters of the responses (Educational Testing Service, 2014, p. 19).

Remediation
The process of supplementary training provided to human raters who demonstrated unacceptable scoring performance levels over a certain time period.

Response
The performance to be evaluated from test tasks in the following formats: short answer, extended answer, essay, presentation, speech sample, demonstration, or portfolio.

Rubric
The scoring criteria used by raters to evaluate responses which include the key response elements and descriptors associated with responses at different levels along the score scale that corresponds to different levels of proficiency of the construct to be measured.

Score equating
A statistical process used to align scores between two parallel test forms, such that identical reported scores from each form hold the same meaning.

Scoring engine

A software architecture that houses all of the necessary routines and components to provide feedback and/or scores for a set of constructed responses.

Scoring leader

An experienced human rater in a supervisory role who is responsible for monitoring the scoring performance and accuracy of raters assigned to them during an operational scoring session.

Scoring rubric

See ***Rubric***.

Sensitivity

The awareness of the contributions of various groups to a society. A sensitivity review of a test also checks to make sure that “there is not the use of stereotyping and language, symbols, words or examples that are sexist, racist or otherwise offensive, inappropriate or negative toward any group.” (ETS, n.d.).

Subgroup

A part of a larger population that is defined on the basis of a characteristic such as gender, race or ethnic origin, training or formal preparation, geographic location, income level, disability, or age (Educational Testing Service, 2014, p. 59).

Task

See ***Constructed-response task***.

Task analysis

The study of a constructed-response task or item to reveal the construct-relevant and construct irrelevant knowledge, skills, and abilities necessary to be able to respond to that task or item.

Threshold

A numerical value for a particular statistic or metric used to support decision-making during the evaluation of a scoring approach.

Use context

The context in which constructed-response scores will be used in decision-making and the intended inferences and actions about individuals which may also relate to possible unintended and negative consequences.

Validity

The extent to which the interpretations of scores, the claims made about test takers, and inferences and actions made on the basis of a set of scores are appropriate and justified by evidence. Validity refers to how the scores are used rather than to the test itself. Validity is a unified concept, but several aspects of validity evidence are often distinguished (Educational Testing Service, 2014, p. 63).

Validity argument
A coherent and rational compilation of evidence designed to convey all of the relevant information available to a program concerning the validity of a test's scores for a particular purpose (Educational Testing Service, 2014, p. 63).

Validity response
A response to a task selected to serve as a sample in order to monitor the operational performance of human raters by comparing raters' scores on the sample to the consensus score assigned by a committee of experts.

www.ets.org