Research Report

Evaluations of Automated Scoring Systems in Practice

ETS RR–20-10

Ourania Rotou
André A. Rupp

December 2020

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

Evaluations of Automated Scoring Systems in Practice

Ourania Rotou & André A. Rupp

Educational Testing Service, Princeton, NJ

This research report provides a description of the processes of evaluating the "deployability" of automated scoring (AS) systems from the perspective of large-scale educational assessments in operational settings. It discusses a comprehensive psychometric evaluation that entails analyses that take into consideration the specific purpose of AS, the test design, the quality of human scores, the data collection design needed to train and evaluate the AS model, and the application of statistics and evaluation criteria. Finally, it notes that an effective evaluation of an AS system requires professional judgment coupled with statistical and psychometric knowledge and understanding of the risk assessment and business metrics.

Keywords Automated scoring; constructed responses; educational assessments; human scoring; model evaluation; natural language processing

doi:10.1002/ets2.12293

The design, implementation, and monitoring of automated scoring (AS) systems within broader constructed response (CR) scoring systems that involve human scoring (HS) components is a complicated process with numerous interrelated design decisions that affect its stability, efficiency, and defensibility (e.g., Bejar, 2011; Habermehl et al., 2020; Rupp, 2018; Schneider & Boyer, 2020; Shaw et al., 2020). Advances in AS systems are driven by scientists in a variety of feeder disciplines such as natural language processing (NLP), speech science, cognitive science, computer science, human–computer interaction, learning analytics, and machine learning to name but a few. The associated disciplinary traditions shape the ways interdisciplinary teams engage with one another to design, evaluate, and implement AS systems and CR scoring processes within a given assessment or learning systems context.

Moreover, AS systems do not function in isolation. Instead, they are embedded within assessment and learning systems, which are developed strategically within a business context and institutions that have their unique missions, priorities, workflows, and technological systems. It is in this complex ecosystem that interdisciplinary teams—which are comprised of members who bring their own disciplinary experiences, expertise, and mindsets to this work—have to function. For example, cognitive scientists are prototypically driven by identifying generalizable scientific findings whereas human–computer interaction specialists may be more comfortable maximizing the affordances of a unique delivery environment where generalizability extends first and foremost to the target population. Similarly, psychometricians are commonly charged with maximizing score quality in terms of reliability, validity, and fairness for particular stakeholder groups. Business managers are typically charged with maximizing revenue, reducing internal cost, and creating attractive products that find relatively broad acceptance with clients. These goals can sometimes be nicely aligned but may also yield conflicts that require interdisciplinary resolution.

In this report, we describe the processes of evaluating the "deployability" of AS systems from the perspective of a large-scale assessment enterprise in which relatively rigorous and conservative quality-control standards for operational processes as well as measurement procedures are in place. We describe the key features of these processes, the core challenges that arise for their management, and the major lessons learned that can help inform others who want to embark on similar endeavors. Conversations with colleagues who work at different institutions have shown that certain details of this process differ across institutions due to the nature of underlying computational systems, external reporting goals, or institutional histories, but many of the associated principles appear to be common.

Thus, we share this information as an illustrative case study rather than to convey prescriptive rules that should be implemented across the board in diverse use contexts. In the end, each organization will have to find its own way

Corresponding author: O. Rotou, E-mail: orotou@ets.org

to institutionalize, codify, and manage these processes within their own ecosystem. Note that we have deliberately excluded detailed discussions of the development of NLP features (e.g., Cahill & Evanini, in press), the idiosyncracies of automated speech recognition (e.g., Zechner & Loukina, 2020), and the emergent field of multimodal analytics (e.g., D'Mello, in press; Khan & Yuchi, 2020) due to the fact that this would require a notable expansion of our discussion.

We have structured this report as follows. In the System Development section, we describe and graphically illustrate the overall multiphase process with a primary focus on the evaluation of the AS system. That is, we touch on the importance of evaluating HS data, but we do not go into detail about how these could be obtained in the first place (see Pedley-Ricker et al., 2020; Wolfe, 2020). In the Evaluation Designs section, we discuss various layers of the evaluation design for AS systems to tease out the contributing design choices that — often implicitly — undergird a particular set of analyses. In the AS Evaluation: Key Statistics section, we focus even more narrowly on the key evaluation statistics and their mindful use using evaluation thresholds. We close the report with comments about overall evaluation mindsets and the procedural complexity of this work.

System Development

In this section, we describe an overall process of AS systems development while focusing most strongly on the evaluation component of the system and on contexts in which AS systems are deployed for the first time in a new use context. Furthermore, we do this from the perspective of a predominantly linear or sequential development cycle in which models are built, evaluated, and then deployed in one cycle with notable gaps between update cycles and relatively small cyclical iterations in places. In more fluid adaptive systems, some of these processes are automated and naturally interwoven (e.g., scoring models may be updated dynamically as sufficient HS with acceptable properties become available); see Habermehl et al. (2020) for an example. Nevertheless, many underlying design and evaluation considerations remain comparable in principle, and human users still have to be able to understand, encode, and review them. Furthermore, they need to be able to communicate qualitative considerations about what constitutes acceptable performance and how associated risks informs system, evaluation, and implementation designs.

Overview

Figure 1 shows a high-level overview of the general workflow for the overall process of designing an AS model evaluation, implementing the evaluation plans, and then taking concrete action on either implementing the model or planning future evaluation studies.

Oversimplifying a bit, an opportunity space for AS is typically identified at the outset by business partners who typically aim to reduce operational scoring costs associated with human raters. A key exception to this situation is digitally delivered assessment and learning systems in which the use of AS is necessary for deployment in the first place to enable real-time feedback and routing. Put simply, sometimes the use of AS is a choice, sometimes it is not. After the initial opportunity is identified, an evaluation of the suitability of the use of existing AS systems for capturing key evidence about the targeted constructs needs to be made, which is informed by sources such as reviews of scientific literature on AS coupled with past empirical experiences of the system in question.

If it seems reasonable to pursue an AS solution, then data must be collected in order to build suitable features or, if a mature scoring engine with an associated feature repertoire already exists, to build scoring models. Depending on the operational context this may require the collection of preoperational data from field trials — often a common occurrence when a new assessment is designed — or the use of operational data. However, in the latter case, this may require that additional HS be conducted. Once various candidate scoring models have been built and evaluated, select models can then be deployed operationally and monitored over time, which may involve "soft" deployment models in which the amount of HS is gradually decreased over time. In even more adaptable systems, the scoring models are updated frequently whenever prediction performance increases further and the entire ecosystem of scoring becomes fully integrated. In the following sections, we provide more detailed descriptions of each phase.

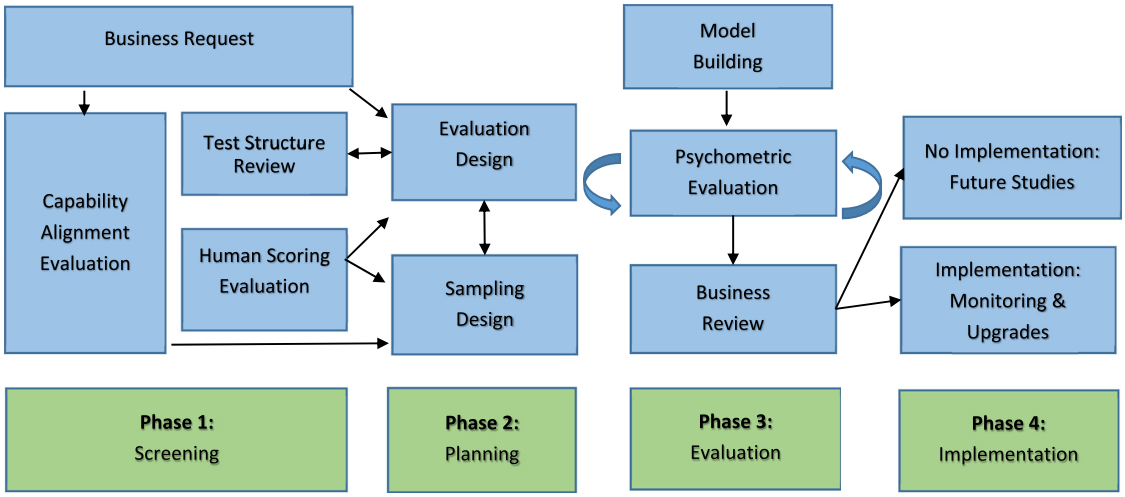

Figure 1 Prototypical workflow/process overview for AS model building and evaluation.

Phase 1: General Screening

The purpose of this phase is to ensure that all stakeholders have a clear and shared understanding of the business objectives for using an AS system in an operational setting. This process often begins with a formal request from either an internal client (e.g., a business unit) or an external client, which is then reviewed by the research team that works on AS solutions.

Specifically, NLP scientists need to determine whether the AS system includes features or routines that can capture key aspects of the targeted constructs. Colleagues who work on assessment/test development often provide information about the target construct, including its operationalization through the HS rubric and associated training materials (e.g., benchmark responses) to help the NLP scientists make judgments about construct alignment. This work is critical not merely for novel assessments that measure novel constructs (e.g., critical reasoning, collaborative problem-solving, emotional dispositions) but also for novel task types within existing assessments that were designed to capture novel aspects of an existing construct (e.g., coherence of discourse in addition to linguistic accuracy, strategy type in addition to misconceptions/skills). In some instances, the NLP and assessment design specialists may need to modify and or adjust the scoring rubric that would make the item more desirable candidate for machine scoring.

Beyond design information for individual items or tasks, it is also important to obtain detailed information on the overall assessment design, which provides insight into the relative weight that scores from automatically scored tasks receive. For example, AS may be used in a high-stakes assessment context as (a) a "check" score (i.e., to monitor HS without using AS for reporting) or as a "contributory" score (i.e., to combine HS and AS for reporting), (b) to identify and filter unusual responses (e.g., cheating/gaming, disengagement), or (c) to route test takers to different activities within an adaptive system based on skill-based performance. In this context, expected score turnaround time is another critical factor because AS systems, the higher-level computational architectures that they are embedded in, and the quality-control processes required to ensure trustworthy scores may not be able to provide scores in a timely manner. It is important to understand all these considerations since they critically affect the scope of work, the associated timelines, the statistical metrics, the evaluation criteria, and the data that are required.

In addition, it is important to understand how scores are reported to stakeholders (e.g., what kinds of supplementary information about score meaning and disclaimers are provided), what kinds of stakes are associated with various potential score uses (e.g., high-stakes admission decisions, broader group-level comparisons), and what kinds of degrees of freedom the stakeholders have to influence the stakes around any use (e.g., automatic required reporting, self-reporting). In our experience, many "contentious" decisions about the use of AS in operational practice have their roots in multifaceted validation and fairness arguments (see, e.g., Bennett & Zhang, 2016) and the consequences that decisions based on reported scores have for learners, teachers, or other score users. Put simply, the higher the stakes are, the more conservative the evaluation guidelines and practices generally are, and vice versa; there is no single "demarcation line" existing between these scenarios despite the publication of landmark papers such as Williamson et al. (2012).

Another critical aspect of AS concerns the quality (e.g., degree of construct representation) and consistency (i.e., intra- and interrater reliability) of HS because they are often conceptually viewed — and statistically treated — as the "gold standard" for model building and evaluation. If HS quality is poor — as reflected in, for example, low interrater agreement or intrarater accuracy as measured by prescored responses that had expert-derived consensus scores — no amount of AS design can compensate. Similarly, if score distributions have undesirable properties, such as ceiling or floor effects or restrictions of range, these could pose additional problems for the AS system.

After the review of the AS system capabilities and the evaluation of HS quality, it typically becomes necessary to make a go or no-go decision as to whether the overall process should continue; if stopped, should resources be invested to improve the AS capabilities or HS quality first. For example, evidence of poor HS quality would suggest that modifications to the training, calibrating, monitoring, and remediation of human raters are advisable before further resources are invested into AS modeling. This information can then be used by business unit management to adjust and finalize the business objective(s) and terms of services.

All of the work in this phase is critical to ensure that services are not overpromised to clients (e.g., promising solutions for types of responses that cannot be reliably scored, promising comprehensive construct coverage when it can only be partially covered, underestimating development and evaluation efforts) and to ensure that possible alternative uses of AS that could still be beneficial to clients are not overlooked even when primary uses of AS seem out of reach (e.g., identification of responses with special characteristics, selection of stimuli with particular characteristics).

Phase 2: Evaluation Planning

The work in this phase is concerned with the creation of data collection and analysis plans for successful model building and evaluation; we describe the latter in more detail in the next section. We note briefly here that the complexity of this work can be influenced notably by the choice of modeling approach. For example, consider the differences between a tree-based or neural-network classification approach with lower-level linguistic features (e.g., *n*-grams, counts of errors) versus a multiple linear regression approach with higher-level aggregate features (e.g., grammatical errors, syntactic complexity, discourse coherence). These approaches require very different sample sizes, corpora, and model fine-tuning to demonstrate appropriate model-data fit and to collect empirical evidence for construct representation and "semantic explainability" of predictions for score users.

The steps for this phase are typically not performed in a sequential order but, rather, in parallel. If it is decided to proceed with the AS model building and evaluation, then sampling rules, variables for model building and evaluation, and analysis plans with their associated statistical metrics and evaluation thresholds need to be determined. This work is often best done by a team that consists of data scientists and psychometricians with their respective emphases on training and experience. For example, data scientists are highly skilled at evaluating the conditions under which different machine learning algorithms perform whereas psychometricians are highly skilled at evaluating how information from such algorithms can be aggregated reliably into complex psychometric latent-variable models. In more automated systems, these decisions are not made sequentially and consciously at fixed time points, but the underlying rationales, decision-making rules, and associated values of the analysts for what is acceptable are encoded in the scripts that drive these processes.

Phase 3: Evaluation Implementation

This phase includes the processes of model building and evaluation. This work is best done by data scientists and psychometricians in consultation with colleagues from NLP and test development when it comes to decision-making processes, which could potentially be formalized through technical advisory committee (TAC) evaluations or other governing bodies. Although it is critical to carefully execute the detailed data-analysis plans developed previously, it is important to recognize that there are always places in which human judgment remains and in which there will be some qualitative judgment based on considerations of acceptability of different risks that can never be fully removed. At some point in this process, it is typically important to formally summarize the available evidence as well as the overall reasoning regarding the factors that support deployment and the factors that pose additional risks. In fact, finding language to articulate what kinds of risks may exist (e.g., score misinterpretation and misuse, hidden financial costs, unfair treatment of population subgroups) and how to quantify these effectively to guide decision-making is often an Achilles heel of the process.

This documentation is often done in technical reports but also includes the use of shorter memoranda that can be more easily digested by decision-makers at higher level of management as well as colleagues who are not quite as technically trained (e.g., colleagues in test development or in the business unit). This kind of documentation can be a helpful political mechanism for "unearthing" unspoken assumptions, beliefs, and values of those decision-makers. As a corollary, it is equally important to capture the resulting decisions and rationales of higher levels of management as an institutional record because institutions typically look for internal and external precedents as benchmarks. This documentation helps to avoid paradoxical situations in which numerous scientific papers or technical reports might be available but no clear documentation around the actual deployment decision can be extracted; see Schneider and Boyer (2020) as well as Shaw et al. (2020) for a similar emphasis on documentation.

To put this another way: despite the availability of various deployment models in practice, deployment of any model at any given point in time—automated or not—is always a go or no-go and, thus, binary decision. In many cases, evaluation results reside in the "messy middle" for AS systems in that some aspects may be highly satisfactory (e.g., prediction of narrow construct aspects is excellent), some aspects may be just good enough (e.g., HS quality meets minimum standards but does not exceed them by much), and some aspects are somewhat problematic (e.g., some construct aspects are captured very well while others are not captured at all or captured only partially). In automated systems, the mapping of this conceptual messy middle onto binary rules is hardwired into the utilized scripts and routines, which carries its own risk (i.e., not being given sufficient attention as key assessment factors change).

Phase 4: The Implementation Phase

Connected to the previous point, the implementation phase involves the creation of procedural guidelines and quality-control protocols as well as the fine-tuning of the computational systems for operational usage. Critical evaluation components include the selection or development of

- Statistical metrics and thresholds for monitoring purposes,
- Plans for the frequency of model or engine upgrades,
- Plans for updating detectors for aberrant response/gaming behavior,
- Data-collection plans for HS for the purpose of building future models and ongoing monitoring,
- Score adjudication rules (if needed),
- Rescoring rules for cases that were scored using AS and HS, and
- Scoring rules for responses that cannot be scored with machines.

It may, of course, be the case that the AS models are not performing in a satisfactory manner in which case implementation cannot be recommended and either more HS data need to be collected or more development work on the AS engine needs to be completed first before reconsideration of deployment.

Evaluation Designs

In this section, we provide a critical discussion of common aspects of evaluation designs for CR scoring solutions and the embedded AS systems. As before, our discussion is informed by a predominantly linear deployment approach in a high-stakes test context rather than a fully adaptive system, but discussions with colleagues have shown us that design and reasoning principles, as well as some of the statistical tools, are commonplace.

Design Aspects

It is helpful to think of the evaluation of a CR scoring approach and its embedded AS system like the process of designing other kinds of experimental research studies that require rigorous and explicit designs (see, e.g., Kirk, 2013). In the particular context of AS, the following four key evaluation design aspects must be considered jointly when determining appropriate quantitative analyses:

- Sampling and partitioning
- Score aggregation

- Score computation
- Score transformation

Furthermore, it is quite common to have multiple data sets from multiple data-collection efforts available to serve different related analytical needs. In the following, we provide a brief overview of these four evaluation design aspects.

Aspect 1: Sampling and Partitioning

This design aspect refers to the ways analysts go about obtaining and managing data for building and evaluating AS models. A few common samples and associated sampling partition scenarios are:

- No sampling = current population data available
- Preoperational sample = sample collected under potentially nonoperational conditions
- Operational sample = sample collected under operational conditions
- Reliability (sub)sample = (sub)sample with multiple human scores for each response
- Repeater (sub)sample = (sub)sample with repeated observations from the same learner
- Feature development (sub)sample = (sub)sample used to build and evaluate new features
    - Feature build (FB) partition = used to build new scoring features
    - Feature evaluation (FE) partition = used to evaluate new scoring features
- Model development (sub)sample = (sub)sample used to build and evaluate new AS models
    - Model build (MB) partition = used to train an AS model
    - Model evaluation (ME) partition = used to evaluate an AS model
- Scoring model development (sub)sample = (sub)sample used to build and evaluate a scoring model with HS and AS components
    - Scoring model build (SMB) partition = used to build new scoring models
    - Scoring model evaluation (SME) partition = used to evaluate new scoring models
- Custom samples and partitions

For example, if population data are available then there is no uncertainty about population patterns and associated statistical parameters, and inferential procedures are not needed. The distinction between a preoperational and an operational (sub)sample matters for considerations about generalizability. Specifically, AS models built from a preoperational sample may not necessarily perform comparably under operational conditions given that motivation levels of learners, administration conditions, and other factors that can affect true score distributions might vary across the two settings. Reliability and repeater (sub)samples may come from special studies or from operational samples and can be used to answer very specific research questions about the consistency of scores.

The next three pairs of (sub)samples (i.e., feature development, model development, and scoring model development) connect to three distinct goals for the development and evaluation of scoring approaches. Generally speaking, independent samples and associated partitions are preferred by analysts so as to avoid confounding of interpretations and decisions across different stages of development. We have encountered situations in which one (sub)sample was needed for NLP research to develop novel construct-centered scoring features, another (sub)sample was needed for statistical research to develop an AS model using HS as the main criterion/outcome variable, and yet another (sub)sample was needed for statistical research to develop a joint scoring approach in which both AS and HS serve as predictors for scale scores (see Breyer et al., 2017; Haberman & Qian, 2014). In each of these three use cases, the (sub)sample needs to be ideally further divided into two independent training and an evaluation partitions or, alternatively, cross-validation approaches with subsequent aggregation of performance statistics across folds need to be utilized for smaller sample sizes (e.g., Hastie et al., 2009, Report 7).

Aspect 2: Score Aggregation

This design aspect refers to the levels within the data at which different kinds of analyses can be conducted. The choice of which of all possible combinations of analyses and levels need to be conducted, evaluated, and utilized for decision-

making—human guided, semiautomated, or fully automated—depends on the goals of the analysis and its associated evidentiary requirements. A few common levels for analysis are as follows:

- Stimulus
  - Task
  - Prompt
- Structure
  - Section
  - Test
  - Administration/form
  - Trait/dimension
- Subgroup
- Rater
- Custom levels

For example, the distinction between task- and prompt-level analyses matters most for so-called generic scoring models that are trained and deployed on a collection of prompts designed to be exchangeable; when so-called prompt-specific scoring models are used, the evaluation of each model at the item/task and prompt level are equivalent. The distinction between section- and test-level analyses can be relevant for score reporting analyses since each can provide different types of evidence about how constructs are measured and related to one another. Administration-level analyses can be equivalent to form-level analyses but are called out here because they almost always involve the utilization of multiple forms and may require the use of cross-form or cross-administration analyses with longitudinal data-analysis components (e.g., Lee & Haberman, 2013). Trait/dimension-level analyses are relevant for multidimensional scoring rubrics, in particular within a single content domain, and may require tools from the area of confirmatory factor analysis or structural equation modeling (e.g., Kline, 2011).

Subgroup analyses are always desirable as one pathway for exploring fairness in scoring, although limited sample sizes within a given (sub)sample or partition may prevent analysts from doing so accurately. Rater-level analyses can provide insight into understanding key performance differences within a human rater population along such dimensions as leniency, centrality, and accuracy (e.g., Nieto & Casabianca, 2019, as well as the entire special issue that this article is in). Finally, we included a custom level as a placeholder for a variety of analyses that may be uniquely needed for a particular application such as evaluations across experimentally manipulated task types, administration conditions, or delivery formats for more traditional assessments or time sequences, activity clusters, or levels in adaptive learning systems.

Aspect 3: Score Computation

The third design aspect concerns the variety of ways in which scores can be created to perform analyses at different aggregation levels within different samples or partitions. Each analysis at each score level may utilize one or multiple sets of scores, depending on whether absolute or relative comparisons of scoring models, learners, or tasks are desired (to name but a few), the structure of the associated statistics, and what are the suitable reference points for relative comparisons (e.g., human true scores vs. reported section scores). A few common scores involved in score computations for AS systems are as follows:

- Human Score 1
- Human Score 2
- Weighted combination of two (or more) human scores
- One machine score
- Weighted combination of one (or more) human scores with one (or more) machine scores
- Weighted combination of two (or more) machine scores without any human scores
- Custom computations

For example, some AS models are trained on a single HS while others are trained on two or more HS. If multiple HS are naturally available for each response, then they can be used to compute interrater agreement statistics without

incurring extra data-collection efforts. This occurrence is a relatively rare in real practice, however, and randomly selected, double-scored consistency/reliability samples are typically needed when only single HS is used for reporting.

Having double-scored human ratings available is generally preferred on statistical grounds. If second human ratings are only collected in cases for which the first HS and the AS are notably different and require adjudication, then agreement statistics will tend to underestimate human – human agreement overall. Moreover, even if double-scored and adjudicated score are available, only the randomly double-scored data should be used in this situation without including scores from cases that required additional HS during adjudication.

The number of HS available affects decisions about rounding in the computation of evaluation statistics if averages are used for training or evaluation of models. In operational score reporting for assessments, three use cases are quite common: (a) reporting a weighted combination of (at least one) AS with (at least one) HS, (b) reporting a single AS without HS, and (c) reporting a weighted combination of multiple AS without HS. The latter two cases are often the desired use cases from a business perspective since they result in the largest amount of operational cost savings even if HS are typically still collected for a small percentage of cases for ongoing quality-control monitoring.

Historically, as we briefly mentioned earlier, AS have also sometimes been used as mere quality-control mechanisms for human raters (i.e., behind the scenes) in a check-scoring approach to determine when additional HS for a response are needed (i.e., whenever the machine and the first human rater disagree notably). However, that deployment model appears to be far less common nowadays because it does not realize as many cost savings as the other deployment models noted above, even though it can serve as a supplemental objective monitoring approach if its construct coverage limitations are acknowledged.

Aspect 4: Score Transformation

The fourth design aspect concerns the way scores are treated before models are built and evaluated. Each evaluation analysis may utilize one or several score transformations that are applied to one or multiple sets of these scores. A few common score transformations are as follows:

- No transformation/raw scores
- Truncation
- Rounding
- Scaling
- Linking/equating
- Custom transformations

For example, if there is a choice between using a rounded and an unrounded score, the unrounded version is always preferred statistically based on the principle that no statistical information is lost in the process. However, if rounded scores are reported to stakeholders, then the evaluation should include computation with rounded scores. Although one can expect a distortion of the key methodological effects at some transformation levels when rounded scores are used (e.g., certain biases may be masked due to rounding or scaling) and should use the obtained insight only in a supplementary manner, analyses with rounded scores can be important as a "sanity check" for reporting evaluations. A related principle pertains to the use of scaling, linking, and equating procedures, which should be done at the last possible step in the evaluation that is concerned with reported scores rather than when an individual item-level scoring model is built.

Section Summary

Table 1 shows how some key design choices may come together in a particular set of analyses, which is shown here for purely illustrative purposes and not to suggest a comprehensive analysis plan for a particular application. The terminology of evaluation, extrapolation, and generalization is taken from Williamson et al. (2012) and refers to different score and interpretation properties.

The broader point within this section is to be aware of the particular design choices that are made for a given evaluation and how these design choices, coupled with the chosen evaluation statistics, affect the veracity of interpretations viz-a-viz the claims that they support. In the next section, we briefly review the psychometric properties of common evaluation statistics.

Table 1 Illustrations of evaluation designs

Analysis type	Analysis focus	Sample	Level	Scores	Weight	Transformation	Selected statistics
Evaluation	H – H agreement	Reliability	Item	Human 1, Human 2	N/A	H unrounded	QWK
Evaluation	H – M agreement	Reliability [ME partition]	Section	Human 1, Human 2 Human 1, Machine	H/H = .50/.50 H/M = .75/.25	H unrounded, M unrounded	QWK, SMD, degradation
Evaluation	H – M agreement	Operational [ME partition]	Item	Human 1, Machine	N/A	H unrounded, M unrounded	Pearson's *r*, QWK, SMD
Evaluation	Score precision	Operational [ME partition]	Test	Human 1, Machine	H/M = .75/.25	H unrounded, M unrounded, unscaled	PRMSE, MSE, RMSE
Extrapolation	Score association	Operational [ME partition]	Section	Human 1, Machine	H/H = .50/.50 H/M = .75/.25	H unrounded, M unrounded, combined score scaled	Pearson's *r*
Generalization	Score association	Operational [ME partition]	Section	Human1, Human 2 Human 1, Machine	H/H = .50/.50 H/M = .75/.25	H unrounded, M unrounded, combined score scaled	Pearson's *r*

Note. H = human; M = machine; MB = model building; ME = model evaluation; MSE = mean squared error; PRMSE = proportional reduction in mean squared error; QWK = quadratic-weighted Kappa; RMSE = root mean squared error; SMD = standardized mean score difference.

Table 2 Properties of evaluation statistics

Statistic	Focus of statistic	Score types	Assumptions	Sampling distribution	Common thresholds
% (exact) agreement	Association	Rounded	Score independence	Empirical only	Depends on scale
K, LWK, QWK	Association	Rounded	Score independence	Theoretical, empirical	.70
SMD	Central tendency	Rounded, real-valued	Depends on formula	Theoretical, empirical	.10, .15
MSE, RMSE	Predictive accuracy	Rounded, real-valued	Depends on model	Theoretical, empirical	Depends on scale
PRMSE	Predictive accuracy	Rounded, real-valued	Depends on model	Empirical only	Depends on scale

Note. K = Kappa; LWK = linear weighted Kappa; QWK = quadratic weighted Kappa; SMD =; MSE = mean squared error; RMSE = root mean squared error; PRMSE = proportional reduction in mean squared error. All statistics can be used for binary and polytomous score variables so no distinction between the two is made here. Common thresholds are listed here for reference and should not be used without further critical examination in new use contexts.

AS Evaluation: Key Statistics

We start our review with a quick disclaimer, namely that the set of evaluation statistics in this report is subject to modification over time as the scientific state-of-the-art in statistics and psychometrics advances and as certain statistics become obsolete, require modifications, or are added to the toolkit. Put differently, best practices are subject to continual reevaluation and adjustment to remain current. We have provided a brief summary of the key considerations around these statistics in Table 2.

We also note briefly that it is generally helpful to plot and summarize score (or feature) distributions at a given design level via common mechanisms such as distributional plots (e.g., histograms, box-and-whiskers plots, scatterplots, density

plots), moment-based statistics (e.g., mean, standard deviation, variance, skewness, kurtosis), quantile-based statistics (e.g., median, interquartile range, other relevant percentiles), counts of observations (e.g., by score point conditional on HS or AS), and pass/fail rates (when applicable) to name but a few. We do not provide additional information on these summaries because they are common statistical practice.

In addition, we refer to application-specific historical benchmarks and associated theoretical expectations for a given use context in the following to help make judgments about appropriateness, but we acknowledge that determining suitable use contexts can, in fact, be a real challenge given the various dimensions along which many assessments vary in their design, implementation, and use.

Measures of Agreement

The following agreement measures assess the alignment between two sets of scores and can be computed for contingency tables of varying sizes. Inspecting these contingency tables can often be helpful for identifying bias in scoring that certain statistics may not be sensitive to; many of the properties discussed here can also be found in handbooks (e.g., Gwet, 2014). Importantly, all measures of agreement are appropriate to use for situations where the pairs of variables can be treated as exchangeable, which is notably the case for interrater agreement evaluations using human ratings. However, they are not truly appropriate for the comparison of nonexchangeable quantities such as human ratings and machine predictions, especially when the scoring models that yielded the predictions were trained on the human ratings in the first place. For these kinds of evaluations, the predictive accuracy measures in the next sections are preferable.

Percentage of Exact and Adjacent Agreement

Percentages of exact and adjacent agreement are simple indices that represent the degree of decision consistency between two independent raters. As their names imply, exact agreement and adjacent agreement are often defined as the percentages of responses that received the same score or the same score plus or minus one score category by two independent raters, respectively. The latter is only meaningful when there are three or more score categories because it is always 100% for binary scores. Both statistics range from 0% to 100% where 0% implies no exact or adjacent agreement—a value never seen in practice; whereas 100% implies perfect exact or adjacent agreement—a value often seen for adjacent agreement but not quite as often for exact agreement.

These two statistics can technically be used to evaluate the agreement between two human raters in a reliability sample (i.e., a sample in which all responses are double scored), which can be done at various score levels. Moreover, they are often requested by clients of AS system providers because they are easy to compute and seemingly easy to interpret. However, from a psychometric perspective, they can result in misleading interpretations due to (a) their scale dependency, (b) their lack of correction for chance agreement, and (c) the tendency of human raters to score responses with points in the middle of the scale. It is thus recommended that these indices should not be used to guide formal evaluation decisions and certainly not as sole criteria. Consequently, there are no commonly agreed-upon thresholds for what constitutes minimally acceptable performance. At a minimum, application-specific historical benchmarks and associated theoretical expectations for a given use context should further inform analyses.

Cohen's Kappa (K) Statistics

Cohen's K measures the percentage of exact agreement but adjusts the computation for chance agreement, which is statistically a very desirable adjustment so as to not overinterpret large values due to the underlying overinflation of base rates for uncorrected statistics. That being said, K statistics are influenced by the distributions of the underlying variables and their relationship, among other factors (e.g., Sim & Wright, 2005). Cohen's K ranges from 0 to 1 where $K = 1$ indicates perfect agreement and $K = 0$ indicates no agreement. A value of .70 could be used as an evaluation threshold for minimally acceptable performance because it represents about 50% shared variance between the two sets of scores akin to other correlational measures. However, application-specific historical benchmarks and associated theoretical expectations for a given use context should further inform analyses.

Generally speaking, any weighted K is an extension of the K statistic, and common variants include linearly weighted K and quadratic-weighted kappa (QWK); as the name implies, the latter includes a nonlinear (i.e., quadratic) weighting

that penalizes larger disagreements more strongly than smaller disagreements. Consequently, K, linearly weighted K, and QWK are identical when there are only two score categories. A common misconception of these statistics is that they are designed only for integer-valued variables. However, formulas exist that extend their definitions to one or two real-valued variables, which can avoid unnecessary rounding.

As with the exact and adjacent agreement statistics, Cohen's K statistics can be used to evaluate the agreement between two human raters in a reliability sample, which can be done at various score levels. It can be shown that if the means and variances of the population distributions for the two variables in question are identical and the two variables are independent, then QWK is identical to the Pearson correlation coefficient, which we review next (see Schuster, 2004).

Pearson Product–Moment Correlation Coefficient (ρ)

The ρ statistic is a common correlational measure in statistics and captures the extent that the scores from two variables are linearly associated. Pearson's ρ ranges from -1 to $+1$ where $\rho = 1$ indicates perfect linear agreement with an increasing association, $\rho = -1$ indicates perfect linear agreement with a decreasing association, and $\rho = 0$ indicates no consistent linear association. Similar to K and QWK, a value of .70 could be used as an evaluation threshold for minimally acceptable performance because it represents about 50% shared variance between the two sets of scores akin to other correlational measures. However, application-specific historical benchmarks and associated theoretical expectations for a given use context should further inform analyses.

Computing estimates of Pearson's ρ for the cases when one variable is real-valued and the other is polytomous or dichotomous results in so-called point-polyserial and point-biserial correlations, respectively. Similarly, if one were willing to make a distributional assumption for the latent distributions underlying the polytomous and dichotomous variables, one could compute so-called polychoric and tetrachoric correlations, respectively. Importantly, the value of Pearson's ρ is not affected by differences in distributional means.

Pearson correlations are often used to compute human–machine agreement at various score levels. They can also be used to examine relationships between reported scores, which are often (at least approximately) real-valued; correlations are commonly used between reported scores across sections that focus on different content domains to support validation arguments. Pearson correlations are also often used to evaluate the degree of linear association between real-valued AS and HS at various levels, which could be done using polychoric or tetrachoric variants that make additional distributional assumptions as noted above. Moreover, if the trend between two variables is monotonically increasing or decreasing in a nonlinear fashion, then Spearman's rank-order coefficient is a superior alternative for quantifying that association.

Standardized Mean Difference

A standardized mean difference (SMD), also commonly known as Cohen's d, measures the similarity of the centers (i.e., means) of two score distributions and is rescaled to account for the variation of these distributions using a pooled variance (e.g., Durlak, 2009). SMD statistics are designed for real-valued score variables even though they can technically be computed for discrete variables as well. They can range from $-\infty$ to $+\infty$ with values closer to 0 indicating very similar distributional means relative to the variation that is in the sample. Unlike for agreement statistics, there is no singular benchmark value that can be recommended as a reasonable statistically driven cutoff despite the fact that values such as .10 or .15 have been previously published. As with other metrics, application-specific historical benchmarks and associated theoretical expectations for a given use context should inform analyses.

The computation of the pooled variance in the denominator depends on the data-collection design; specifically, whether the two sets of ratings are independent or dependent on one another. Similar to the distinction between an independent-samples t-test and a dependent-samples t-test, the computational difference will affect the resulting values of this statistic. For example, two sets of human ratings are commonly independent—and furthermore discrete—while human and machine scores are commonly dependent—with only one being discrete—suggesting two different computation methods for these two different comparisons. Importantly, for subgroup comparisons, it may be advisable to use the pooled standard deviation from the full sample rather than just the subgroup sample.

Measures of Score Accuracy

The measures in this subsection concern the predictive accuracy of scores, rather than the agreement even though these two concepts are related.

Mean Squared Error and Root Mean Squared Error

The mean squared error (MSE) statistic is a common evaluation statistic for linear models because it is a measure of the variation of the prediction error of a model; specifically, it reflects the average squared difference between the predicted and observed values. Due to the squaring of the units, the MSE statistic is not on the same scale as the involved scores, and thus, the square root is often taken resulting in the root mean squared error (RMSE), similar to a general standard deviation measure. The MSE statistic can range from $-\infty$ to $+\infty$ with values closer to 0 indicating a very good fit of the statistical prediction model to the training data.

As with other statistics, application-specific historical benchmarks should further inform the development of appropriate guidelines, and because these measures depend on the scale of the variables, no single threshold can be identified across use contexts. Therefore, absolute evaluations of MSE by itself are difficult to make unless MSE is used as a component in a statistic such as the coefficient of determination discussed below, but relative comparisons across different prediction models to evaluate their relative model-data fit can be very helpful.

Proportional Reduction in Mean Squared Error

The proportional reduction in mean squared error (PRMSE) statistic is a measure of prediction accuracy after adjusting for score unreliability, which shares some structural similarities with reliability coefficients under certain circumstances. The PRMSE statistic incorporates the inherent value of AS relative to HS in a single statistic within a predictive, rather than a relative comparison, framework.

Similar to MSE, the PRMSE statistic is computed in square units. The PRMSE statistic ranges from 0 to 1 with PRMSE = 1 indicating perfect prediction. Similar to other reliability statistics, values of at least .80 or .90 could be used as an evaluation threshold for minimally acceptable performance but application-specific historical benchmarks and associated theoretical expectations for a given use context should further inform analyses.

Measures of Degradation

Degradation statistics are not new sets of statistics but, rather, difference statistics that are used for comparative analyses under two different conditions; a common comparison is that of human–machine agreement relative to human–human agreement. Technically, degradation variants of all statistics discussed above could be computed, but differences between QWK, Pearson's ρ, and SMDs across human–human and human–machine conditions are most common. The range of degradation statistics depends on the range of values of the ingredient statistics, but values close to 0 always indicate very similar values of the ingredient statistic under the two computation conditions. As before, application-specific historical benchmarks and associated theoretical expectations should further inform the development of appropriate guidelines.

Sampling Distributions

A key aspect of statistical inference is concerned with making judgments from a sample of data back to a reference distribution via hypothesis testing and confidence interval construction that make use of the standard errors of these statistics. Within a frequentist estimation framework, this judgment requires that an exact sampling distribution is known, that a normal-theory approximation can be used, or that empirical sampling distributions are constructed via mechanisms such as bootstrapping or jack-knifing. Within a Bayesian estimation framework, this judgment requires the use of posterior distribution and associated test statistics.

In some AS contexts, the overall sample size may be large enough to effectively nullify the value of inferential procedures because the point estimates are essentially error-free. However, this may not always be the case, especially when smaller preoperational samples are available or when analyses are done at a design level that effectively partitions the data into

smaller sets (e.g., subgroup analyses). For many clients of AS system providers, inferential statistics are often difficult to interpret; nevertheless, they can be very useful for internal evaluation purposes.

Concluding Thoughts

We conclude this report with two overarching sets of comments: one on the issue of evaluation mindset and one on the issue of procedural complexities.

Evaluation Mindset

Determining the psychometric quality of the scores that come from an AS system is a vital component for fair and valid assessments. From a construct-centered perspective, key quality considerations center around the ability of scores and scoring routines to help capture essential aspects of the target construct so that fair and valid decisions about stakeholders can be made through the assessment overall. In other words, we do not want the AS system to compromise the overall integrity of the assessment design and, ideally, would like to use it to leverage its complementary capabilities relative to HS. From a purely statistical perspective, key quality considerations center around the reliability of AS, their ability to represent true HS whenever those are used as a gold-standard criterion, and understanding the trickle-down effect of using AS for individual items or tasks on reported scale scores for different reporting levels.

As we discussed in this report, a comprehensive psychometric evaluation entails analyses that take into consideration the specific purpose of AS, the test design, the quality of HS, the data collection design needed to train and evaluate the AS model, and the application of statistics and evaluation criteria, all finally cumulating in a decision about the implementation of AS in an operational setting. In addition, decisions need to be made about the ongoing monitoring of the AS system, which has implications for updating AS models, filtering rules, routing routines, or other components of such systems (see Rupp, 2018; Schneider & Boyer, 2020; Shaw et al., 2020).

An effective evaluation of an AS system requires professional judgment coupled with statistical knowledge and empirical historical experience in a blend of scientific rigor and artful practice, a point that we have reiterated frequently in this report. In addition, we want to reiterate a few aspects tied to risk assessment, simplistic thinking, data properties, statistical goals, baseline selection, and business metrics.

Risk Assessment

Various implications come down to the simple issue of proper "risk assessment" for the deployment of an AS system, often in the context of validity and fairness tied to score reports or diagnostic feedback and its use. As we noted earlier, evaluating the different kinds of risks and identifying suitable metrics that communicate such risks are often critical in the evaluation enterprise. Typically, more stringent criteria for acceptability of an AS model may need to be adopted in situations when AS is the sole score or a contributing score with high weights in a high-stakes assessment context compared to when AS is used only for internal routing purposes or as a contributory score in a lower-stakes context. Perhaps most importantly, broader guidelines always need to be fine-tuned and adapted based on the use case information for each AS application. These guidelines apply to the design of the analyses as well as to the statistics and decision-making thresholds that are utilized while realizing that not all analyses may be technically possible for every AS application.

Simplistic Thinking

Evaluation thresholds can be useful for decision-making in that they can seemingly simplify the evaluation task. However, rather than subscribing to any single number, it is much more important to use a mixture of scientific insight about the general statistical principles (e.g., using unrounded, real-valued numbers increases computational accuracy; scaling prematurely unnecessarily degrades performance), statistical properties of a particular statistic (e.g., a value of a correlation coefficient of .70 reflects approximately 50% of shared variance between the two involved variables), historical data from comparable contexts (e.g., from a previous engine version, deployment window, or related investigation), and stakes-based qualitative considerations (e.g., applying more lenient criteria for less critical roles of AS) in decision-making. Importantly, there is a clear risk in blindly applying seemingly accepted guidelines based on a few peer-reviewed publications in the area rather than making smart principled decisions that are suitable to the use context in question.

Data Properties

Computing statistics without properly correcting for key underlying data properties distorts the values of these statistics and all associated interpretations and should not be used. For example, for the case of agreement statistics, chance-corrected measures such as *K* and QWK are much more preferable for decision-making than simpler uncorrected measures such as percentage (adjacent) agreement even if they are easier to communicate to some audiences. Similarly, the computation of SMD critically depends on the proper computation of the pooled standard deviation in the denominator and inappropriate formulas can lead to misinterpretations of score differences, especially if multiple variants of SMDs coexist in the same technical report.

Modeling Goals

Agreement statistics are generally preferable to quantify human – human agreement whereas predictive accuracy measures are generally preferable to quantify the performance of an AS model. Consequently, if this principle was followed strictly, it would not allow for the direct comparison of human – human and human – machine agreement using a common metric such as QWK. However, it would allow one to evaluate the relative performance of each independently in order to question the statistical integrity of the scoring approach overall.

Baseline Selection

For all statistics the selection of a reference point or baseline is really important because it affects the resulting interpretations. For example, if the analytic focus is on how much information is lost when using an AS in a weighted combination with an HS relative to using two HS, one might compute an SMD with the double HS as the baseline. However, at a lower item level, it may be that the AS is compared directly to a single HS. If an operational scoring model is a candidate for replacement by an updated scoring model, the old model may become the baseline. Such baselines could, potentially, also come from historical data, external sources, or theoretical expectations. For example, if an AS model is designed to capture construct-relevant information that is mostly complementary to what human raters can evaluate, then having a low human – machine agreement may actually be desirable.

Business Metrics

From a business perspective, there are often additional metrics that have value, which includes all metrics that can capture any potential cost savings. For example, hypothetically, if it were internally known that the scoring of a particular response by one human rater costs about $5 at an annual volume of 3 million responses, then one can estimate that the elimination of, say, even just 10% of these ratings could save $1.5 million. Yet caution is needed in such computations because there are typically notable, indirect start-up costs for the development, integration, and maintenance of the overall system, so these would have to be counterbalanced against the seemingly direct cost savings. This kind of systemic thinking was nicely demonstrated in a recent paper from the area of automated item generation (Kosh et al., 2019).

Procedural Complexities

As we have reiterated in this report and as illustrated in the flowchart in Figure 1, designing, implementing, and monitoring a CR scoring approach is a highly interdisciplinary endeavor and involves the coordination of various teams with different internalized experiences, practices, and mindsets. Public discussions around CR scoring generally — as well as specifically — typically revolve first and foremost around broader issues of validity, fairness, and privacy. Such discussions are commonly framed in the context of the use goals and associated theory of action and change for the assessment or learning environment into which CR scoring approaches are embedded rather than the AS systems per se.

Comprehensive and thoughtful responses to how these issues are addressed require an intellectual and empirical "crosswalk" of the evidence that is gathered through the interdisciplinary process sketched out in this report. This crosswalk has to be integrated with a parallel — and partially interwoven — crosswalk of evidence from supplementary validation studies. These studies may use some of the CR scores (e.g., for correlational analyses) but typically also supplement them with qualitative evidence from interviews, surveys, cognitive laboratories, and other experimental setups. Although this evidence

may be partially acknowledged or even utilized in some of the design and development work for AS and HS approaches, there are typically also validation questions that transcend those narrower needs. In other words, a comprehensive answer to validity and fairness issues requires broad, well-organized "portfolio of evidence" in which computational results from CR scoring approaches play but one critical role.

In this report, we were only able to scratch the surface of the complexity of these issues and other sources can be consulted for some complementary perspectives. For comprehensive overviews of different sources of validity evidence around AS systems we refer the reader to landmark references like Williamson et al. (2012) and Bennett and Zhang (2016). For comprehensive overviews of different aspects of AS system development in operational practice one should consider the recently published handbook by Yan et al. (2020). As additional reference points for AS evaluation principles readers may consult the joint *Standards* for *Educational and Psychological Testing* (American Educational Research Association et al., 2014), the *Standards for Quality and Fairness* (Educational Testing Service, 2014), the *ITC Guidelines on Test Use* (International Test Commission, 2000), the *Standards Catalogue* (ISO/IEC/IEEE, 2017), or other relevant disciplinary standards documents (see Haisfield et al., 2018).

In closing, AS development is a resource-intensive, complex, and challenging process, perhaps even more so as novel technologies in areas such as speech processing, multimodal analytics, and learning analytics/machine learning as well as associated digital learning and assessment systems proliferate. Such environments are effectively novel instruments that afford us novel perspectives on how learners engage, reason, and self-regulate with the activities we design for them. The complexity of the required evidence and arguments for validity and fairness will only increase, especially as colleagues all over the world seek to transform how we think about learning and assessments. Being principled, transparent, and thoughtful will remain as important as ever, which is why it is important not to view the information in this report as prescriptive but, rather, as illustrative. We hope that it can help inform decision-making processes in interdisciplinary teams that are working in similar contexts and that it encourages others who work in different contexts — and who might disagree with some of the statements we made — to share openly their approaches and underlying rationales.

References

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

Bejar, I. I. (2011). A validity-based approach to quality-control and assurance of automated scoring. *Assessment in Education: Principles, Policy & Practice, 18*(3), 319 – 341. https://doi.org/10.1080/0969594X.2011.555329

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

Breyer, J. F., Rupp, A. A., & Bridgeman, B. (2017). *Implementing a contributory scoring approach for the GRE Analytical Writing section: A comprehensive empirical investigation* (Research Report RR-17-14). Educational Testing Service.

Cahill, A., & Evanini, K. (in press). Natural language processing for writing and speaking. In D. Yan, A. A. Rupp, & P. Foltz (Eds.), *Handbook of automated scoring: Theory into practice*. Routledge/CRC Press.

D'Mello, S. (in press). Multimodal analytics. In D. Yan, A. A. Rupp, & P. Foltz (Eds.), *Handbook of automated scoring: Theory into practice*. Routledge/CRC Press.

Durlak, J. (2009). How to select, calculate, and interpret effect sizes. *Journal of Pediatric Psychology*, *34*(9), 917 – 928. https://doi.org/10.1093/jpepsy/jsp004

Educational Testing Service. (2014). *ETS standards for quality and fairness*. https://www.ets.org/s/about/pdf/standards.pdf

Gwet, K. L. (2014). *Handbook of interrater reliability* (4th ed.). Advanced Analytics. https://doi.org/10.1002/9781118445112.stat06882

Haberman, S., & Qian, J. (2014). *The best linear predictor for true score from a direct estimate and several derived estimates* (Research Report No. RR-04-35). Educational Testing Service.

Habermehl, K., Nagarajan, A., & Dooley, S. (2020). A seamless integration of human and automated scoring. In D. Yan, A. A. Rupp, & P. Foltz (Eds.), *Handbook of automated scoring: Theory into practice*. CRC Press.

Haisfield, L., Yao, E., & Wood, S. (2018, April 12 – 16). *Industry standards for an emerging technology: Automated scoring* [Paper presentation]. National Council on Measurement in Education Annual Meeting, New York, NY, United States.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning* (2nd ed.). Springer. https://doi.org/10.1007/978-0-387-84858-7

International Test Commission (ITC) (2000). *International guidelines for test use*. www.intestcom.org

ISO, ITC, & IEEE. (2017). *Systems and softward engineering - Requirements for testers and reviewers of information for users*. https://www.iso.org/obp/ui/#iso:std:iso-iec-ieee:26513:ed-1:v1:en

Khan, S. M., & Yuchi, H. (2020). Deep learning networks for automated scoring applications. In D. Yan, A. A. Rupp, & P. Foltz (Eds.), *Handbook of automated scoring: Theory into practice*. Routledge/CRC Press.

Kirk, G. (2013). *Experimental design: Procedures for the behavioral sciences* (4th ed.). Sage. https://doi.org/10.4135/9781483384733

Kline, R. (2011). *Principles and practice of structural equation modeling* (3rd ed.). Guilford Press.

Kosh, A. E., Simpson, M. A., Bickel, L., Kellogg, M., & Sanford-Moore, E. (2019). A cost–benefit analysis of automatic item generation. *Educational Measurement: Issues and Practice, 38*(1), 48–53. https://doi.org/10.1111/emip.12237

Lee, Y.-H., & Haberman, S. J. (2013). Harmonic regression and scale stability. *Psychometrika, 78*(4), 815–829. https://doi.org/10.1007/s11336-013-9337-1

Nieto, R., & Casabianca, J. M. (2019). Accounting for rater effects with the hierarchical rater model framework when scoring simple structured constructed response items. *Journal of Educational Measurement, 56*, 547–581. https://doi.org/10.1111/jedm.12225

Pedley-Ricker, K. L., Hines, S., & Connelly, C. (2020). Operational human scoring at scale. In D. Yan, A. A. Rupp, & P. Foltz (Eds.), *Handbook of automated scoring: Theory into practice*. Routledge/CRC Press.

Rupp, A. A. (2018). Designing, evaluating, and deploying automated scoring systems with validity in mind: Methodological design decisions. *Applied Measurement in Education, 31*(3), 191–214. https://doi.org/10.1080/08957347.2018.1464448

Schneider, C., & Boyer, M. (2020). Design, development, and implementation of automated scoring systems. In D. Yan, A. A. Rupp, & P. Foltz (Eds.), *Handbook of automated scoring: Theory into practice*. Routledge/CRC Press.

Schuster, C. (2004). A note on the interpretation of weighted kappa and its relation to other rater agreement statistics for metric scales. *Educational and Psychological Measurement, 64*(2), 243–253. https://doi.org/10.1177/0013164403260197

Shaw, D., Bolender, B., & Meisner, R. (2020). Quality-control for automated scoring in large-scale assessment. In D. Yan, A. A. Rupp, & P. Foltz (Eds.), *Handbook of automated scoring: Theory into practice*. Routledge/CRC Press.

Sim, J., & Wright, C. C. (2005). The kappa statistic in reliability studies: Use, interpretation, and sample size requirements. *Physical Therapy, 85*(3), 257–268. https://doi.org/10.1093/ptj/85.3.257

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

Wolfe, E. (2020). Human scoring with automated scoring in mind. In D. Yan, A. A. Rupp, & P. Foltz (Eds.), *Handbook of automated scoring: Theory into practice*. Routledge/CRC Press.

Yan, D., Rupp, A. A., & Foltz, P. W. (2020). *Handbook of automated scoring: Theory into practice*. Routledge/CRC Press.

Zechner, K., & Loukina, A. (2020). Automated scoring of extended spontaneous speech. In D. Yan, A. A. Rupp, & P. Foltz (Eds.), *Handbook of automated scoring: Theory into practice*. Routledge/CRC Press.

Suggested citation:

Rotou, O., & Rupp, A. A. (2020). *Evaluations of automated scoring systems in practice* (Research Report No. RR-20-10). Educational Testing Service. https://doi.org/10.1002/ets2.12293

Action Editor: James Carlson

Reviewers: Issac Bejar and Kathryn Pedley