Establishing Standards of Best Practice in Automated Scoring

Scott Wood, Erin Yao, Lisa Haisfield, and Susan Lottridge

Introduction

Many professions have published standards of best practices for their members to follow, which provide direction to professionals about how best to conduct their work. Some standards include ethical guidelines to ensure that professionals conduct their business in ways that are fair and considerate toward others. Most importantly, published standards provide confidence to stakeholders, including customers, who are invested in the work that professionals offer, by documenting expectations of best practice and ethical professional behavior.

For professionals in the assessment industry, the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014) are frequently cited as standards of best practice and ethical behavior. For professionals involved in large-scale assessment at the state level, the *Operational Best Practices for Statewide Large-Scale Assessment Programs* provide additional guidelines (Council of Chief State School Officers [CCSSO] & Association of Test Publishers [ATP], 2013). The International Test Commission maintains six sets of published guidelines for assessment: *The ITC Guidelines on Adapting Tests* (International Test Commission, 2017), *The ITC Guidelines on Test Use* (International Test Commission, 2001), *The ITC Guidelines on Computer-Based and Internet-delivered Testing* (International Test Commission, 2006), *The ITC Guidelines on Quality Control in Scoring, Test Analysis and Reporting of Test Scores* (International Test Commission, 2014a), *The ITC Guidelines on the Security of Tests, Examinations, and Other Assessments* (International Test Commission, 2014b), and *The ITC Guidelines on Practitioner Use of Test Revisions, Obsolete Tests, and Test Disposal* (International Test Commission, 2015).

Professions in fields that overlap with educational assessment may have their own standards of best practice. For example, statisticians have the *Ethical Guidelines for Statistical Practice* (American Statistical Association, 2018). Data scientists have the

Data Science Code of Professional Conduct (Data Science Association, n.d.). Linguists have the *Linguistic Society of America Code of Ethics* (Linguistic Society of America, 2019). Educational researchers have the *American Educational Research Association Code of Ethics* (American Educational Research Association, 2011). Computer programmers have both the *Association of Computing Machinery Code of Ethics and Professional Conduct* (Association for Computing Machinery, 2018) and the *Association of Computing Machinery/International of Electrical and Electronics Engineers Computer Society Software Engineering Code of Ethics and Professional Practice* (Gotterbarn, Miller, & Rogerson, 1997).

For assessment professionals who are also automated scoring (AS) professionals, there is no single set of standards of best practice. Certainly, standards can be extracted from the documents cited above, as AS professionals are involved with assessment, statistics, data science, linguistics, educational research, and computer programming. Additional standards can be extracted from seminal papers, books, and technical reports in the field of AS. However, these standards are not centralized in one location for AS professionals and AS consumers to utilize.

This paper reviews the assessment and AS literature to identify key standards of best practice and ethical behavior for AS professionals and codifies those standards in a single resource. Having a unified set of AS standards is important for several reasons. First, given that AS is an emerging technology in educational assessment, it is important to establish guidelines of good practice for professionals and stakeholders learning to use this new technology. Second, due to the wide variety of professionals involved in AS technology development (e.g., psychometricians, linguists, data scientists, and computer programmers), a unified set of standards would guide these diverse professionals toward common objectives. Third, and most importantly, having standards for which stakeholders can hold AS professionals accountable can provide stakeholders with greater confidence in the use of AS.

The next section describes the methods used to identify, review, and summarize the standards and recommendations in the AS literature. This is followed by a summary of 10 important, high-level standards for AS professionals. Each standard is supported by exemplar citations from the AS literature. A summary and references list appear at the end of the report.

Methods

Source Identification

AS research staff searched for documents that referenced AS standards, processes, recommendations, or implementations. The set of sources was not expected to be

exhaustive of the literature; rather, it was intended to identify those references illustrative of best practice.

The sources can be divided into a few categories:

- Standards, guidelines, or recommendations from professional organizations
- Published work offering frameworks on the use and evaluation of AS
- Results of large-scale programs that use standards to evaluate AS
- Published work offering frameworks on the evaluation of machine learning models

In all, 16 sources were identified as representative of the literature. Table 1 presents the 16 sources, organized by the four categories above. Five sources represented standards, guidelines, or recommendations from professional organizations. Six sources were published works offering frameworks on the evaluation of AS. Four sources cited results and evaluation standards on large-scale programs, including national programs. One source provided evaluation standards used in machine learning model evaluations—a set of evaluations broader than those relevant to AS. This last source was included because it provided an excellent overview of machine learning evaluation methods for the novice user of machine learning.

Table 1. Sources Used for Standards Review

Standards, Guidelines, or Recommendations from Professional Organizations
American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
Council of Chief State School Officers, & Association of Test Publishers. (2013). Scoring. In *Operational best practices for statewide large-scale assessment programs*. (pp. 125–134) Washington, DC: Council of Chief State School Officers and the Association of Test Publishers.
International Test Commission. (2014a). ITC guidelines on quality control in scoring, test analysis, and reporting of test scores. *International Journal of Testing*, *14*(3), 195–217.
Joint Committee on Testing Practices. (2004). *Code of fair testing practices in education*. Washington, DC: Joint Committee on Testing Practices. *https://www.apa.org/science/programs/testing/fair-testing.pdf*.
International Test Commission. (2006). International guidelines on computer-based and internet-delivered testing. *International Journal of Testing*, *6*(2), 143–171.

Published Work Offering Frameworks on the Use and Evaluation of Automated Scoring
Madnani, N., Loukina, A., von Davier, A., Burstein, J., & Cahill, A. (2017, April). Building better open-source tools to support fairness in automated scoring. In D. Hovy, S. Spruit, M. Mitchell, E. Bender, M. Strube, & H. Wallach (Eds.), *Proceedings of the first ACL workshop on ethics in natural language processing* (pp. 41–52).
Powers, D. E., Escoffery, D. S., & Duchnowski, M. P. (2015). Validating automated essay scoring: A (modest) refinement of the "gold standard." *Applied Measurement in Education, 28*(2), 130–142.
Shermis, M. D., Burstein, J., Elliot, N., Miel, S., & Foltz, P. (2016). Automated writing evaluation: An expanding body of knowledge. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (2nd ed., pp. 395–409). New York, NY: Guilford.
Williamson, D. M., Bennett, R. E., Lazer, S., Bernstein, J., Foltz, P., Landauer, T. K., Rubin, D. P., Way, W. D., & Sweeney, K. (2010). *Automated scoring for the assessment of Common Core Standards*. ETS, Pearson, & The College Board.
Yang, Y., Buckendahl, C. W., Juszkiewicz, P. J., & Bhola, D. S. (2002). A review of strategies for validating computer-automated scoring. *Applied Measurement in Education, 15*(4), 391–412.
Yang, Y., Buckendahl, C. W., Juszkiewicz, P. J., & Bhola, D. S. (n.d.). *Validating computer automated scoring: A conceptual framework and a review of strategies* [Unpublished manuscript].
Results of Large-Scale Programs on the Use of Standards to Evaluate and Monitor Automated Scoring
McGraw-Hill Education CTB (2014). *Smarter Balanced Assessment Consortium--Field test: Automated scoring research studies*. Monterey, CA: Smarter Balanced Assessment Consortium.
Pearson, & Educational Testing Service. (2015). *Research results of PARCC automated scoring proof of concept study*.
Wang, Z., & von Davier, A. A. (2014). *Monitoring of scoring using the e-rater automated scoring system and human raters on a writing test* (ETS Research Report ETS RR-14-04). Princeton, NJ: ETS.
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.
Published Work Offering Frameworks on the Evaluation of Machine Learning Models
Zheng, A. (2015). *Evaluating machine learning models: A beginner's guide to key concepts and pitfalls*. Sebastopol, CA: O'Reilly Media.

Source Review and Summarization

AS research staff divided the sources amongst themselves for further reading and standard identification. Staff members took notes where sources described a

standard or best practice. Notes were aggregated across the reviewers. Commonalities across the notes were grouped together. This process resulted in 10 core high-level AS standards of best practice. All authors reviewed the 10 core standards and refined the language based on experience with AS projects.

Results

This section presents the key standards of best practice and ethical behavior for AS professionals, as identified by the authors of this report.

Standard 1: Automated scores should achieve industry absolute and relative thresholds for accuracy when compared with human scores.

Numerous publications have identified AS and human scoring metrics suitable for evaluating the reliability and accuracy of AS scoring. Such metrics can be evaluated via *absolute* or *relative* thresholds. Absolute thresholds are used when a metric is compared to a constant value, such as when a human-AS quadratic weighted kappa is compared to 0.70. Relative thresholds are used when a metric is compared to a corresponding human or human-human metric for the item, such as when a human-AS quadratic weighted kappa is compared to the human-human metric.

AS agreement with a human rater is frequently cited in the literature. At a minimum, desired levels of agreement with human raters at the summed score and trait level must be demonstrated. These statistics demonstrate consistency in scoring with an expert human rater using statistics such as exact agreement, quadratic weighted kappa, and standardized mean differences (Williamson, Xi, & Breyer, 2012).

Table 2 includes several common metrics used to evaluate human-AS accuracy, including the standardized mean difference, the standard deviation (SD) ratio, the difference in exact agreement rates, quadratic weighted kappa, and the difference in quadratic weighted kappa. Included are thresholds recommended by AS professionals. Models meeting these thresholds are considered appropriate for operational use.

Table 2. Common Metrics for Evaluating AS Accuracy, with Thresholds

Metric	Threshold	Reference
Standardized Mean Difference between Human and AS	$-0.15 \leq SMD \leq 0.15$	Williamson, Xi, & Breyer (2012)
Standard Deviation (SD) Ratio between Human and AS	$2/3 \leq (SD_{Human} / SD_{AS}) \leq 1.50$	Wang & von Davier (2014)
Difference between Human-Human and Human-AS Exact Agreement Rate	$EA_{Human,Human} - EA_{Human,AS} \leq 5.125\%$	McGraw-Hill Education CTB (2014); Pearson & ETS (2014)
Quadratic Weighted Kappa between Human and AS	$QWK_{Human,AS} \geq 0.70$	Williamson, Xi, & Breyer (2012)
Difference between Human-Human and Human-AS Quadratic Weighted Kappa	$QWK_{Human,Human} - QWK_{Human,AS} \leq 0.10$	Williamson, Xi, & Breyer (2012)

Table 3 provides exemplar citations from the literature review that support this standard.

Table 3. Exemplar Citations Supporting Standard 1

Source	Examples
AERA, APA, & NCME, 2014	"Standard 2.7: When subjective judgement enters into test scoring, evidence should be provided on both interrater consistency in scoring and within-examinee consistency over repeated measurements. A clear distinction should be made among reliability data based on (a) independent panels of raters scoring the same performances or products, (b) a single panel scoring successive performances or new products, and (c) independent panels scoring successive performances or new products" (p. 44).
	"[Standard 2.7] Comment: Task to task variation in the quality of an examinee's performance and rater to rater consistencies in scoring represent independent sources of measurement error. Reports of reliability/precision studies should make clear which of these sources are reflected in the data. Generalizability studies and variance component analyses can be helpful in estimating the error variances arising from each source of error. These analyses can provide separate error variance estimates for tasks, for judges, and for occasions within the time period of trait stability. Information should be provided on the qualifications and trainings of the judges used in the reliability studies. Interrater or interobserver agreement may be particularly important for ratings and observational data that involve subtle discriminations. It should be noted, however, that when raters evaluate positively correlated characteristics, a favorable or unfavorable assessment of one trait might color their opinions of other traits. Moreover, high interrater consistency does not imply high examinee consistency from task to task. Therefore, interrater agreement does not guarantee high reliability of examinee scores" (p. 44).
	"Standard 2.8: When constructed-response tests are scored locally, reliability/precision data should be gathered and reported for the local scoring when adequate-size samples are available" (p. 44).
Williamson, Xi, & Breyer, 2012	"The model building and evaluation process for automated scoring is largely dependent on the quality of human scores…[I]f the inter-rater agreement of independent human raters is low, especially below the .70 threshold, then automated scoring is disadvantaged in demonstrating this level of performance…" (p. 7).
	"In typical practice at ETS, we first conduct the empirical associations with human score (agreement, degradation, and standardized mean score difference) at the task level. At the task type level (aggregated results across the individual tasks within the task type) and the reported section score level the entire contingent of measures discussed above are also employed in the evaluation of performance" (p. 8).

	"Empirical Performance: Associated with Typical Scoring Method (Human Scores). This entails making sure [1] Human scoring process and core quality …, [2] Agreement of automated scores with human scores" using QWK (.70), Pearson correlations (.70), human rater reliability among other raters (so H1-H2), "[3] Degradation from the human-human score agreement" AES-Human cannot be more than .1 lower than H1-H2 to ensure that .70 QWK doesn't allow a pass for if the AES model is deficient as compared to the H1-H2 reliability, "[4] Standardized mean score difference between human and automated scores" SMD between human and automated scores cannot exceed .15, "[5] Threshold for human adjudication" how much difference is required before adjudication is used, "[6] Human intervention of automated scoring" response characteristics that render AES inappropriate for scoring, and "[7] Evaluation at the task type and reported score level" look at distribution of changes in reported scores that would results from AES at the task score level (p. 7).
	"The model building and evaluation process for automated scoring is largely dependent on the quality of human scores…[I]f the inter-rater agreement of independent human raters is low, especially below the .70 threshold, then automated scoring is disadvantaged in demonstrating this level of performance…" (p. 7).
CCSSO & ATP, 2013	"When AI engines are used in a place of human scoring, or for confirmation or quality control of human scoring, scoring procedures should meet the same standards for accuracy and reliability that exist for human scoring of the same item type" (p. 131).
Yang et al., 2002	"…[A] straightforward way to demonstrate the accuracy and appropriateness of CAS [computer-automated scoring] -system-generated scores is to evaluate their relationship to the scores assigned by human scorers to the same item (e.g., task, prompt) or the same scores given by CAS systems have a high level of agreement with those trained scorers….Furthermore, one can also compare agreement between two human experts and between a CAS system and a human expert to demonstrate that a CAS system is no less consistent than human experts" (p. 400).
	"…[O]ne can investigate the reliability of CAS-system-generated scores by correlating them with expert scores as well as by comparing the reliability of scores assigned by human scorers and by a CAS system" (p. 400).
	"It is also possible to approximate the true scores by using the consensus scores given by a group of experts…These consensus scores…are the scores a group of experts agreed on after discussion" (p. 401). "If a CAS system produces scores that agree completely with a human rater, it may indicate that the system not only modeled the construct-relevant aspects of a scoring process but also possibly emulated the personal and situational characteristics that may contribute to the errors and biases in measurement" (p. 401).

Williamson et al., 2010	"Automated scores are consistent with the scores from expert human grader" (p. 5). The distribution of scores should approximate human scores of essays.
	"This similarity is typically demonstrated through statistical measures of agreement between automated and human scores, such as correlations and weighted kappa (rather than percent agreement, which may overestimate the agreement rate between automated and human scores)" (p. 5).
Shermis et al., 2016	"Individual-response-level measures included exact agreement, exact+adjacent agreement, kappa, quadratic weighted kappa, and the Pearson product moment correlation" (p. 402).
Pearson & ETS, 2015	"Evaluation criteria for the scoring models was based on criteria most often used in evaluating automated scoring…and consisted of the following measures of inter-rater agreement: Pearson correlation, quadratic-weighted kappa, exact and adjacent agreement, and standardized mean difference" (p. 9-10). The resource also provides summed score metrics and by trait metrics, as well as score point distribution metrics, recall, precision, and F1.
	"Throughout the report, we include discussions of the percentage of prompt / trait combinations that might be considered to perform less well using Williamson et al.'s criteria, as a way of comparing human performance to automated scoring performance" (p. 10). This resource discusses the typical metrics between human-machine difference (relative) and absolute threshold overall for Pearson Correlation, QWK [quadratic weighted kappa], EA [exact agreement], and SMD [standardized mean difference] (p. 11).
Yang et al., n.d.	"First, one should perform a test on the similarity of score distributions produced by the raters. This can be done by testing marginal homogeneity of the two raters' scores … If marginal homogeneity is not rejected, Kappa, or preferably, according to Zwick (1988), Scott's π coefficient can be used to further assess chance-corrected agreement" (p. 16)
	"The first series of analyses examined the agreement between the pairwise comparisons of total score distributions. These analyses utilized two non-parametric tests, the Kolmogorov-Smirnov (K-S) two-sample test and the Wilcoxon-Mann-Whitney (W-M-W) test….[S]tatistical significance tests were conducted using an alpha level of .10" (p. 22).

McGraw-Hill Education CTB, 2014	"Bridgeman (2013) noted that the high agreement between two raters can occur when raters are truncating the rubric score range. CTB has found that an engine's quadratic weighted kappa (QWK) may be high even though the engine exact agreement rate in comparison is low. In this situation, engines are usually giving adjacent scores to humans so that both the percent agreement and kappa statistics are not comparable to humans. For this reason, CTB also monitors engine performance for a notable reduction (greater than 0.05 difference) in perfect agreement rates between the human-human and engine-human scores" (p. 15). Statistical criteria can be divided into 3 broad categories: evaluated against the final human scores of records, evaluated against the inter-rater performance of the two initial human raters, and evaluated for the performance in different subgroups.
	"Note the difference between the evaluation criteria in the first and second category. For the first category, the scores assigned by the Automated Scoring system are compared against the final human scores of record. For the second category, statistics from the first category are compared against performance of the two human raters…In other words, evaluation of the criteria of the second category should be subsequent to evaluation of the criteria in the first category. Hence, one could argue that these three categories constitute a hierarchy. For example, if an Automated Scoring system does not meet the performance criteria for the entire population, then evaluating its performance on subgroups may be less relevant" (p. 17).

Standard 2: AS engines and procedures should be transparently described such that stakeholders understand how they operate and whether they satisfy construct coverage.

Transparency about what response features are used in AS is another frequently cited standard. This information is crucial to determine if an AS engine can properly score responses to items designed to elicit evidence of a certain construct (International Test Commission, 2006). Construct coverage is the ability of an AS engine's candidate feature set to reflect the construct being assessed.

For example, in essay scoring, an AS score that primarily depends on word count for scoring would be called into question given this standard (Williamson et al., 2010). Word count does not have a meaningful relationship to the quality of writing or knowledge of content and therefore should not be utilized as the sole determinant of an AS score. In addition, the features used to predict scores may not adequately represent the breadth of the construct, thereby introducing bias.

Transparency in AS goes beyond the candidate feature set used for training the engine. Where possible, all aspects of human and automated scoring should be documented and made available to stakeholders. Such information can include:

- how the data were collected for engine training,
- how human raters produced scores for the training sample,
- how the engine was trained,
- how accurate the scoring models are, and
- how AS and human scoring will be used together for operational scoring.

If both paper and online tests are administered, comparability studies should make it clear when AS will be used. If human scoring will be used for paper testing and AS for online tests, comparability studies are necessary to show that one format will not produce biased scores relative to the other.

Some item types are better suited for automated scoring than others. For example, portfolio assessment and items involving hand-drawn input would be challenging to score accurately via AS. If an assessment program uses AS for a unique item type, AS professionals should be transparent about why AS was chosen as a scoring method, including evidence that AS scores a unique item type accurately.

Table 4 provides exemplar citations from the literature review that support this standard.

Table 4. Exemplar Citations Supporting Standard 2

Source	Examples
ITC, 2006	In a section titled "Ensure knowledge, competence, and appropriate use of CBT [computer-based testing]/Internet testing": "Document the constructs intended to be measured and investigated" (p. 153).
	"Ensure all those involved in test design and development…have sufficient knowledge and competence to develop CBT/Internet tests" (p. 153).
	"Consider the psychometric qualities of the CBT/Internet test." For example, provide "documentation for psychometric properties of the CBT/Internet test" and "ensure that current psychometric standards (test reliability, validity, etc.) apply even though the way in which the tests are developed and delivered may differ" (p. 155).
	One section focuses on ensuring that an online test (CBT) is equivalent to a paper test, and certain aspects might be likened to human scorers: "Provide clear documented evidence of the equivalence between the CBT/Internet test and noncomputer version…have comparable reliabilities…correlate each other at the expected level from the reliability estimates" (p. 156).

Source	Examples
Williamson et al., 2010	"…[T]he most notable limitation is that automated scoring assumes computer test delivery and data capture, which in turn may require an equation editor or graphing interface that students can use comfortably" (p. 2).
AERA, APA, & NCME, 2014	"Standard 1.9: When a validation rests in part on the opinions or decisions of expert judges, observers, or raters, procedures for selecting such experts and for eliciting judgments or ratings should be fully described. The qualifications and experience of the judges should be presented. The description of procedures should include any training and instructions provided, should indicate whether participants reached their decisions independently, and should report the level of agreement reached. If participants interacted with one another or exchanged information, the procedures through which they may have influenced one another should be set forth" (p. 25–26).
	"Standard 4.18: Procedures for scoring and, if relevant, scoring criteria, should be presented by the test developer with sufficient detail and clarity to maximize accuracy of scoring. Instructions for using rating scales or deriving scores obtained by coding, scaling, or classifying constructed responses should be clear. This is especially critical for constructed-response items such as performance tasks, portfolios, and essays" (p. 91).
	"Standard 4.20: The process for selecting, training, qualifying, and monitoring scorers should be specified by the test developer. The training materials, such as the scoring rubrics and examples of test takers' responses that illustrate the levels on the rubric score scale, and the procedures for training scorers should result in a degree of accuracy and agreement among scorers that allows for the scores to be interpreted as originally intended by the test developer. Specifications should also describe processes for assessing scorer consistency and potential drift over time in raters' scoring" (p. 92).
	"Standard 4.21: When test users are responsible for scoring and scoring requires scorer judgement, the test user is responsible for providing adequate training and instruction to scorers and for examining scorer agreement and accuracy. The test developer should document the expected level of scorer agreement and accuracy and should provide as much technical guidance as possible to aid users in satisfying this standard" (p. 92).
	"Standard 6.8: Those responsible for test scoring should establish scoring protocols. Test scoring that involves human judgment should include rubrics, procedures, and criteria for scoring. When scoring of complex responses is done by computer, the accuracy of the algorithm and processes should be documented" (p. 118).

Source	Examples
	"Standard 6.8: Those responsible for test scoring should establish scoring protocols. Test scoring that involves human judgment should include rubrics, procedures, and criteria for scoring. When scoring of complex responses is done by computer, the accuracy of the algorithm and processes should be documented" (p. 118).
	"Standard 6.11: When automatically generated interpretations of test response protocols or test performance are reported, the sources, rationale, and empirical basis for these interpretations should be available, and their limitations should be described" (p. 119).
	"Standard 4.19: When automated algorithms are to be used to score complex examinee responses, characteristics at each score level should be documented along with the theoretical and empirical bases for the use of the algorithms" (p. 92).
CCSSO & ATP, 2013	"Any measure or analysis used to check accuracy and reliability of the scoring process should be made available for the client's review" (p. 130).
Williamson et al., 2010	"In the area of mathematics, the performance of automated scoring systems is typically quite robust when the response format is constrained" (p. 2).
	"The agreement between human raters may be lower than desired. Thus, agreement with human scores may not always be a sufficient accomplishment" (p. 2).
	"Develop items with definitive correct answers that the automated scoring system can verify" (p. 3) Opinion writing is difficult for AES systems to discern.
	"Also, it is often the case that human raters score short content responses at considerably higher agreement rates than they do essay responses, which creates a higher standard for automated scoring methods to attain" (p. 3).
	"The way automated scores are produced is understandable and somewhat meaningful" (p. 5). Constructs are logical. For example, there should not be scores primarily generated from word count because then test takers would just write long essays.
	"However, it would be undesirable to have scores generated primarily from word count because such generation might encourage the student to maximize the number of words at expense of other, more valued aspects of writing" (p. 5).
Madnani et al., 2017	"...construct irrelevant factors includes continuous covariates which are likely to be correlated with the human scores but are not relevant to the construct measured by the test or, if relevant, should not be the main factor behind the model prediction" (p. 5). For example, the length of an essay may be an example of this. It should be only given relative weight for fairness.

Source	Examples
Williamson et al., 2012	"Construct relevance and representation"—steps proposed to evaluate fit between assessment and AES: "[1] construct evaluation" or match between construct and AES capacity, "[2] task design" or fit between features and test task, "[3] scoring rubric" or consistency between features measured by AES and those on scoring rubric, and "[4] reporting goals" or determining if the goal of the reporting is consistent with AES capabilities (p. 6).
	"Each system has at the core of the capability a set of features that are designed to measure the elements of writing that are computer-identifiable and believed to be relevant to the construct of writing, even if they are not directly equivalent to what a human grader might identify in a similar effort" (p. 3).
	"Automated scoring poses some distinctive validity challenges such as the potential to under- or misrepresent the construct of interest, vulnerability to cheating, impact on examinee behavior, and score users' interpretation and use of scores" (p. 4).
Yang et al., 2002	"…[M]any researchers have also stressed the importance of understanding the scoring processes that CAS [computer-automated scoring] systems used" (p. 402). Similar to content validity, this discusses ensuring that the engine is stressing features that are important to good writing.
	"…by analyzing the patterns and nature of disagreement between expert ratings and CAS-system-generated scores, one may identify the difference between human and computer scoring models in terms of the factors that considered and the relative weighting of these factors" (p. 402–403).
Zheng, 2015	Outlines the need for using an independent validation dataset: "The model training process receives training data and produces a model, which is evaluated on validation data" (p. 20). This details cross-validation, held-out validation, and bootstrapping techniques for evaluating validity of models.
	"Optimal hyperparameter settings often differ for different datasets. Therefore they should be tuned for each dataset" (p. 28).
CCSSO & ATP, 2013	"Methods for the calibration of the artificial intelligence scoring engines, and evidence that the engine meets accuracy and reliability standards, should be documented" (p. 131).

Standard 3: Scores produced by AS should demonstrate fairness for all populations.

Related to validity, providing evidence that AS evaluation is fair for persons from diverse groups is another frequently cited standard. A common approach to addressing this standard is to compare human-to-AS engine mean score differences between a majority group and subgroups of interest (Madnani et al., 2017; Williamson, Xi, & Breyer, 2012). Those differences should be similar, so when larger differences are observed for a subgroup of interest, it indicates possible bias. That is, AS might not be

appropriate if mean differences between human raters and AS scores surpass specified thresholds for subgroups compared to a majority group.

AS professionals should also consider the feature set used in an engine. Construct-irrelevant features should be avoided, especially features that may advantage or disadvantage one subgroup over another. Differential feature functioning provides a statistical methodology to identify any problematic features (Penfield, 2016; Zhang, Dorans, Li, & Rupp, 2017).

Table 5 provides exemplar citations from the literature review that support this standard.

Table 5. Exemplar Citations Supporting Standard 3

Source	Examples
Yang et al., 2002	"In the published literature, both agreement/reliability approaches and true score approaches demonstrate desirable performance of CAS systems across various domains and populations" (p. 401).
AERA, APA, & NCME, 2014	"Standard 3.8: When tests require the scoring of constructed responses, test developers and/or users should collect and report evidence of the validity of score interpretations for relevant subgroups in the intended population of test takers for the intended uses of test scores" (p. 66).
	"[Standard 3.8] Comment: For human scoring, scoring procedures should be designed with the intent that the scores reflect the examinee's standing relative to the tested construct(s) and are not influenced by the perceptions and personal predispositions of the scorers. It is essential that adequate training and calibration of the scorers be carried out and monitored throughout the scoring process to support the consistency of scorer's ratings for individuals from relevant subgroups. Where sample sizes permit, the precision and accuracy of scores for relevant subgroups should also be calculated" (p. 66).
	"[Standards 3.8] Comment: Scoring algorithms need to be reviewed for potential sources of bias. The precision of scores and validity of score interpretations resulting from automated scoring should be evaluated for all relevant subgroups of the intended population" (p. 66–67).
CCSSO & ATP, 2013	Raters should be provided with "information on disregarding cues related to disability, English learner status or accommodations that are unrelated to scoring criteria" (p. 128).
	"AI [Artificial Intelligence] validation should represent student responses representative of the entire population of possible student response submission. Validation should include a range of score points, types and styles of writing, and other relevant considerations" (p. 131).

Source	Examples
McGraw-Hill Education CTB, 2014	"Williamson, Xi, and Breyer (2012) flag the SMD if the difference between automated scores and human scores is greater than .15 in absolute value. Similarly, they flag the SMD for a subgroup if the difference between the automated scores and human scores for that subgroup is greater than .10 in absolute value. Because the larger the population SMD value the more likely the subpopulation SMD value will be flagged, CTB reduced the amount of SMD separation tolerated by flagging the population SMD if it exceeds .12 in absolute value" (p. 15).
Williamson et al., 2010	"Automated scores [should be] fair. It is critical that automated scoring be equitable for persons from diverse groups" (p. 5).
Williamson et al., 2012	"We have established a more stringent criterion of performance, setting the flagging criteria at .10, and applied this criterion to all subgroups of interest to identify patterns of systematic differences in the distribution of scores between human scoring and automated scoring for subgroups" (p. 10).
	"…examining differences in the associations between automated and human scores across subgroups at the task, task type, and reported score levels. Major differences by subgroups may indicate problems with the automated scoring model for these subgroups and should be evaluated for potentially undesirable performance with the subgroups in question" (p. 10).
	Investigating "the generalizability of automated scores by subgroup. Substantial differences across subgroups may suggest that the scores are differentially reliable for different groups" (p. 10).
	Examining "differences in the predictive ability of automated scoring by subgroup. … First is to compare an initial human score and the automated score in their ability to predict the score of a second human rater by subgroup. The second type of prediction is comparing the automated and human score ability to predict an external variable of interest by subgroup" (p. 10).
	"subgroup differences should also be investigated in relation to the decisions made based on the scores. This is the most prominent manifestation of group differences" (p. 10).
Madnani et al., 2017	"RSMTool [software for evaluating subgroup differences in automated scoring] considers how well the automated scores agree with the human scores (or another, user-specified gold standard criterion) and whether this agreement is consistent across different groups of test-takers." (p. 5)
	"RSMTool also includes Differential feature functioning (DFF) analysis…This approach compares the mean values of a given feature for test-takers with the same score but belonging to different subgroups" (p. 5).

Standard 4: Convergent and discriminant validity studies should be conducted to establish empirical relationships between AS scores and other constructs.

Test-criterion relationships reflect how well AS scores relate to relevant constructs as measured by an assessment or observable criterion (e.g., educational or job success) external to the assessment of interest. The recommendation to conduct criterion-related validity studies is commonly cited in the literature (e.g., Powers, Escoffery, & Duchnowski, 2015; Shermis, et al., 2016). As with reliability, the expectation is that the relationship between AS scores and the external measure will be similar in magnitude and direction in comparison to the relationship of human scores with the external measure. For example, compared to human scores, one would expect AS scores from a writing test to have similar correlations with the multiple-choice portion of an English language arts exam.

Table 6 provides exemplar citations from the literature review that support this standard.

Table 6. Exemplar Citations Supporting Standard 4

Source	Examples
Powers et al., 2015	"For example, the specific directions given to raters might be varied experimentally: some raters could be instructed to read essays slowly and deliberately, while others would be directed to read more rapidly (and perhaps) superficially. The validity of automated scores would be supported to the extent that they correlate more strongly with the scores given by deliberate readers than by those given by less careful reader" (p. 141).
Williamson et al., 2010	"Automated scores have been validated against external measures in the same way as is done with human scoring….Examples of relevant external criteria include scores on other test sections, grades in relevant academic classes, scores on the same test section on alternate occasions, and scores on specially designed external measures of the construct of interest" (p. 5).
Williamson et al., 2012	"[I]t is of relevance to investigate more than just the consistency with human scores and to also evaluate the patterns of relationship of automated scores, compared to their human counterparts, with external criteria…These independent variables may be scores on other sections of the same test or external variables that measure similar, related, or different constructs" (p. 9).

Source	Examples
	"Within test relationships: Are automated scores related to scores on other sections of the test in similar ways compared to human scores?; External relationships: Are automated scores related to other external measures of interest in similar ways com- pared to human scores?; Relationship at the task type and reported score level: Are the relationships similar at the task type and reported score level? These comparisons should be made both at the task/task type score level and reported score level" (p. 9).
	"How generalizable are the automated scores across tasks and test forms in comparison to human scores? How generalizable are the automated–human combined scores across test forms? A comparison of the generalizability of human and automated scores across tasks and test forms will provide insights into how consistently students perform across tasks and test forms" (p. 9).
	"To what extent do automated, human, and automated–human combined scores on one test form predict human scores on an alternate form? This analysis will reveal whether the use of automated scoring may improve the alternate form reliability of the scores" (p. 10).
Shermis et al., 2016	"Studies have examined the effects of automated evaluation of writing as well as how they generalize to other criterion measures of student performance. … Results [from one study] indicated that the students using automated feedback received higher grades on their summaries, spent more than twice as much time on writing and revising, and managed to retain the skills they learned…. Results [in another study] showed that students receiving feedback improved their summary writing by an overall effect size of d=0.9 compared to control students" (p. 21).

Standard 5: When implementing AS, consideration must be given to contextual factors such as the stakes associated with test performance, item types, and scoring approached that integrate human and AS.

Given known limitations of AS—for example, critics point out that the engine does not "understand" writing the way a human does—there are many considerations for the implementation of AS in a specific program context. This standard dictates consideration of the stakes associated with test performance, item types, and scoring approaches that integrate human and AS.

Often, decisions about how scoring is implemented depend on if the test is a high-stakes assessment (McGraw-Hill Education CTB, 2014; Yang, et al., 2002). Whereas AS may be used as the sole scoring mechanism for low-stakes tests or feedback tools, it may be more appropriate to integrate AS and human scoring for high-stakes

assessments—perhaps using a resolution score if the human score and AS differ substantially.

Table 7 provides exemplar citations from the literature review that support this standard.

Table 7. Exemplar Citations Supporting Standard 5

Source	Examples
Shermis et al., 2016	"Considerations include the following…: construct-based scoring designs; integrated assessments in which both automated scores and human scores serve inter-related roles; strengthen operational human scoring to support modeling of AWE [Automated Writing Evaluation] systems; augmented use of human scores to broaden construct representation; enhanced understanding of human scoring processes; disclosure of scoring approaches; and use of a variety of evidential categories to justify score use" (p. 26).
Williamson et al., 2012	"A rough ordering (from more conservative to more liberal use) of implementations for use of automated scoring is as follows: … Automated quality control of human scoring. The results of a single human score and an automated score are compared. If there is a discrepancy beyond a certain threshold between the two then the response is sent to a second human grader. The reported score is based solely on the human score (either the single human score or the mean of the two human scores). … Automated and human scoring. The score from a single human grader and automated score are averaged or summed to produce the reported score. Responses with score discrepancies beyond a certain threshold are scored by additional human graders. Proposed reporting policies vary, but adjudication procedures have included reporting the average of all scores provided, as well as reporting the average of the two scores in highest agreement, and several variations of these, conditional on the particular distribution of scores involved. … Automated scoring alone. Reporting scores solely from the automated system. This is the most liberal use of automated scoring" (p. 5).
	"The use of automated scoring for high-stakes decisions is subject to a higher burden of both the amount and quality of evidence to support the intended use than for lower-stakes and practice applications. The choice of implementation policies for automated scoring would be influenced by the quantity and quality of evidence supporting the use of automated scoring, the particular task types, testing purpose, test-taker population to which it is applied, and the degree of receptivity of the population of score users to models of implementation" (p. 5).
Yang et al., 2002	"Differences in the level of integration reflect differences in the perceptions of utility and implications stemming from the use of a CAS system" (p. 408). For example, if CAS is viewed as the human scorer then CAS may have a different validity criteria then if viewed as a read-behind.

Source	Examples
McGraw-Hill Education CTB, 2014	"[R]ead and read behind scenarios … can be categorized based on 1. The number of raters (one or two), 2. the type of the first and second rater (human or Automated Scoring system), and 3. the adjudication rule which determines when scores from the first and second rater need to be adjudicated by the third rater: a. adjudicated when the scores of the first and second rater disagree (non-exact) b. adjudicate when the scores of the first and second rater differ by more than 1 score point (non-adjacent)" (p. 40).

Standard 6: During live testing, accuracy and reliability of AS via process monitoring should be made available to the client.

Providing clients with access to scoring process monitoring is another AS standard crucial to ensure accuracy, reliability, and transparency (Wang & von Davier, 2014). Typically, a dashboard is provided to allow clients or a third party a way to audit scoring accuracy and error rates in real time. This provides a way for scoring errors to be immediately identified and addressed during a testing or scoring event.

Reported metrics should include AS score point distributions, human score point distributions, and human-AS agreement statistics, if human scores are available.

Table 8 provides exemplar citations from the literature review that support this standard.

Table 8. Exemplar Citations Supporting Standard 6

Source	Examples
AERA, APA, & NCME, 2014	"Standard 6.9: Those responsible for test scoring should establish and document quality control processes and criteria. Adequate training should be provided. The quality of scoring should be monitored and documented. Any systematic source of scoring errors should be documented and corrected" (p. 118).
CCSSO & ATP, 2013	"Any measure or analysis used to check accuracy and reliability of the scoring process should be made available for the client's review" (p. 131).
ITC, 2014	"Independent Monitoring of Quality Control Procedures… should be carried out in collaboration with all stakeholders, with the aim of auditing specific processes, for example, monitoring inter-rater reliability and checking data entry error rates" (p. 204).

Standard 7: It is essential to evaluate the quality of inputs to an AS engine (responses, human scoring, universe of acceptable responses) before training.

Another standard identified in the review affirms the importance of evaluating the quality of inputs to the AS engine, including item responses, human scores for those responses, and the item's universe of acceptable responses (Williamson, et al., 2010; Williamson, Xi, & Breyer, 2012). First, responses that are considered non-attempts (e.g., blank response, gibberish, refusals, etc.)—as identified by agreed-upon scoring rules—are separated from valid attempts and processed using a different workflow. Specifically, the non-attempt responses are used to establish rules and models for AS to assign condition codes, while valid attempts are used to create the scoring models. Second, AS models developed with data from low-quality human scoring will result in poor AS engine performance. Third, the item's universe of acceptable responses can impact engine performance. For instance, differences in the number of concepts or ways of describing these concepts elicited by the item can affect the suitability of an item for AS. An item such as "Describe the characteristics of the chemical element mercury" might be more suitable for AS than the item "Describe how 19th-century American wars led to the expansion of the United States via manifest destiny." The first item is likely to elicit several standard and common characteristics of mercury (silver colored, liquid at room temperature, poisonous, used in thermometers), while the second item is likely to elicit a broad range of many ideas (War of 1812, Mexican-American War, etc.) and may not therefore be as well suited to AS.

Table 9 provides exemplar citations from the literature review that support this standard.

Table 9. Exemplar Citations Supporting Standard 7

Source	Examples
Williamson et al., 2012	"[T]he model building and evaluation process for automated scoring is largely dependent on the quality of human scores…[I]f the inter-rater agreement of independent human raters is low, especially below the .70 threshold, then automated scoring is disadvantaged in demonstrating this level of performance" (p. 7).

Standard 8: The impact or consequences of AS on the test or reported score should be considered and documented.

Related to the stakes or context of the assessment, it is imperative to clearly define and consider the impact or consequences of AS on the test or reported score (Joint Committee on Testing Practices, 2004; Williamson, Xi, & Breyer, 2012). This standard recommends understanding how the use of AS may affect aspects of the test-taker's experience (e.g., if students write differently given their knowledge that the test is

scored by AS). Additionally, information about AS accuracy of score-based decisions should be transparently communicated, so that these scores are used responsibly at the test and item level. For example, if AS is used for a writing assessment to establish the proficiency of an examinee, AS professionals and test developers should be transparent around potential misclassification errors and proper interpretation of proficiency classifications.

Table 10 provides exemplar citations from the literature review that support this standard.

Table 10. Exemplar Citations Supporting Standard 8

Source	Examples
Williamson et al., 2010	"The impact of automated scoring on reported scores is understood [by AS professionals and stakeholders]" (p. 6).
Williamson et al., 2012	"What impact does the use of automated scoring have on the accuracy of score-based decisions? In some contexts, assessment scores are used for classification purposes, for example, a binary decision about eligibility for admissions or exemption from English language coursework once admitted, or a decision regarding placing students into several levels of English class. Depending on the intended use of the assessment scores, the aggregated reported scores may be subject to further analyses to see if human–machine combined scores introduce a greater amount of decision errors than human scores" (p. 10).
	"What claims and disclosures should be communicated to score users to ensure appropriate use of scores? Researchers should work with the operational program to establish a common understanding of the intended claims and intent for disclosure of both strengths and limitations of automated scoring to ensure an informed population of score users. These claims and disclosures may include the extent to which different aspects of the target construct are covered by automated scoring and its major construct limitations" (p. 10).
	"What consequences will the use of automated scoring bring about? Replacing one human rater with automated scoring or using automated scoring to quality-control human scoring may change users' perceptions of the assessment, how users interpret and use the scores for decision-making, how test takers prepare for the test, and how the relevant knowledge and skills are taught" (p. 10).
	"Automated scoring poses some distinctive validity challenges such as the potential to under- or misrepresent the construct of interest, vulnerability to cheating, impact on examinee behavior, and score users' interpretation and use of scores" (p. 4).
	"What are the response characteristics that render automated scoring inappropriate? … Currently the e-rater technology will flag essays of excessive length or brevity, repetition, those with too many problems, or off-topic responses for scoring by human raters. This adds additional support for the quality of the scores produced" (p. 8).

Standard 9: Procedures should be in place to identify alert papers (i.e., responses reflecting cheating or disturbing content).

A less frequently recognized core standard is that there should be engine procedures to identify "alert papers"—that is, responses reflecting cheating or disturbing content, to which a school might need to respond (Council of Chief State Schools, & Association of Test Publishers, 2013; Williamson, Xi, & Breyer, 2012). Engines typically have mechanisms to identify and flag such responses, such as a filter that detects responses that have been copied from other sources. Hybrid approaches with keyword and machine learning techniques can also be used to flag alert papers.

These mechanisms become crucial if only AS, and not human scoring, is used in operational practice. Cheating detection is important to maintain the integrity and validity of the test scores. Disturbing content can be a liability if an examinee follows through on any threats present in their response. These threats can include harming themselves, others, or property.

Table 11 provides exemplar citations from the literature review that support this standard.

Table 11. Exemplar Citations Supporting Standard 9

Source	Examples
CCSSO & ATP, 2013	"Procedures should be established to identify, to evaluate, and if necessary, escalate alert papers … to the client" (p. 133).

Standard 10: Policies around how and when to recalibrate the engine should be established.

The final core standard is that policies must be established to determine when it is appropriate to recalibrate the engine (Council of Chief State Schools Officers and Association of Test Publishers, 2013). When there are major changes made to the program or population, it is necessary to recalibrate the engine. For example, changing the population of test-takers from 4th graders to 6th graders would warrant recalibration.

Recalibration should also be considered if:

- the score point distribution (using human scores) changes significantly from the score point distribution of human scores used for training, or
- the human-AS agreement rates decrease over time.

Without this type of maintenance of scoring models, models may become less accurate and not meet performance standards and thresholds previously discussed.

Table 12 provides exemplar citations from the literature review that support this standard.

Table 12. Exemplar Citations Supporting Standard 10

Source	Examples
CCSSO & ATP, 2013	"AI performance results should be measured and analyzed regularly. A process should be established to permit recalibration and/or retraining, as appropriate" (p. 131).

Conclusion

This report identified 10 AS standards common to the academic literature and professional standards in assessment, data science, and related fields. Having a unified set of AS standards establishes guidelines of good practice for an emerging technology, guides a diverse group of AS professionals, and provides stakeholders with the confidence that AS professionals are conducting their work in a proficient way.

AS staff at ACT plan to review these standards on an annual basis. This ensures that the standards are based on current best practices and new research findings. Major updates to these standards will be published as needed.

The authors encourage feedback about these standards from AS professionals. The standards are influenced by our experiences working with customers using AS but may represent only a partial view of the industry. AS professionals are invited to contact our team at *CRASE@act.org* to provide feedback and suggestions, which we will incorporate into future versions of these standards.

Finally, the authors hope that this document, along with similar documents being developed at other organizations using AS, can become the foundation for industry-wide standards used worldwide. Such standards should ensure that AS yields accurate and reliable scores that meet the expectations of our stakeholders.

References

American Educational Research Association. (2011). AERA code of ethics. *Educational Researcher, 40*(3), 145–156.

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

American Statistical Association. (2018). *Ethical guidelines for statistical practice*. Alexandria, VA: Committee on Professional Ethics of the American Statistical Association.

Association for Computing Machinery. (2018). *ACM code of ethics and professional conduct*. *https://www.acm.org/code-of-ethics*.

Council of Chief State School Officers, & Association of Test Publishers. (2013). Scoring. In *Operational best practices for statewide large-scale assessment programs*. (pp. 125–134). Washington, DC: Council of Chief State School Officers and the Association of Test Publishers.

Data Science Association. (n.d.). *Data science code of professional conduct*. *https://www.datascienceassn.org/code-of-conduct.html*.

Gotterbarn, D., Miller, K., & Rogerson, S. (1997). Software engineering code of ethics. *Communications of the Association for Computing Machinery, 40*(11), 110–118.

International Test Commission. (2001). International guidelines for test use. *International Journal of Testing, 1*(2), 93–114.

International Test Commission. (2006). International guidelines on computer-based and internet-delivered testing. *International Journal of Testing, 6*(2), 143–171.

International Test Commission. (2014a). ITC guidelines on quality control in scoring, test analysis, and reporting of test scores. *International Journal of Testing, 14*(3), 195–217.

International Test Commission. (2014b). *International guidelines on the security of tests, examinations, and other assessments*. Lincoln, NE: ITC. *https://www.intestcom.org/files/guideline_test_security.pdf*.

International Test Commission. (2015). *International guidelines for practitioner use of test revisions, obsolete tests, and test disposal*. Lincoln, NE: ITC. *https://www.intestcom.org/files/guideline_test_disposal.pdf*.

International Test Commission. (2017). *The ITC guidelines for translating and adapting tests* (2nd Ed.). Lincoln, NE: ITC. *https://www.intestcom.org/files/guideline_test_adaptation_2ed.pdf*.

Joint Committee on Testing Practices. (2004). *Code of fair testing practices in education*. Washington, DC: Joint Committee on Testing Practices. *https://www.apa.org/science/programs/testing/fair-testing.pdf*.

Linguistic Society of America. (2019). *LSA code of ethics*. *https://www.linguisticsociety.org/content/lsa-revised-ethics-statement-approved-july-2019*.

Madnani, N., Loukina, A., von Davier, A., Burstein, J., & Cahill, A. (2017, April). Building better open-source tools to support fairness in automated scoring. In D. Hovy, S. Spruit, M. Mitchell, E. Bender, M. Strube, & H. Wallach (Eds.), *Proceedings of the first ACL workshop on ethics in natural language processing* (pp. 41–52).

McGraw-Hill Education CTB. (2014). *Smarter Balanced Assessment Consortium--Field test: Automated scoring research studies*. Monterey, CA: Smarter Balanced Assessment Consortium.

Pearson, & Educational Testing Service. (2015). *Research results of PARCC automated scoring proof of concept study*.

Penfield, R. (2016). Fairness in test scoring. In N. Dorans & L. Cook (Eds.), *Fairness in educational assessment and measurement* (pp. 55–75). New York, NY: Routledge.

Powers, D. E., Escoffery, D. S., & Duchnowski, M. P. (2015). Validating automated essay scoring: A (modest) refinement of the "gold standard." *Applied Measurement in Education*, *28*(2), 130–142.

Shermis, M. D., Burstein, J., Elliot, N., Miel, S., & Foltz, P. (2016). Automated writing evaluation: An expanding body of knowledge. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (2nd ed., pp. 395–409). New York, NY: Guilford.

Wang, Z., & von Davier, A. A. (2014). *Monitoring of scoring using the e-rater automated scoring system and human raters on a writing test* (ETS Research Report ETS RR-14-04). Princeton, NJ: ETS.

Williamson, D. M., Bennett, R. E., Lazer, S., Bernstein, J., Foltz, P., Landauer, T. K., Rubin, D. P., Way, W. D., & Sweeney, K. (2010). *Automated scoring for the assessment of Common Core Standards*. ETS, Pearson, & The College Board.

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.

Yang, Y., Buckendahl, C. W., Juszkiewicz, P. J., & Bhola, D. S. (2002). A review of strategies for validating computer-automated scoring. *Applied Measurement in Education*, *15*(4), 391–412.

Yang, Y., Buckendahl, C. W., Juszkiewicz, P. J., & Bhola, D. S. (n.d.). *Validating computer automated scoring: A conceptual framework and a review of strategies* [Unpublished manuscript]. The Gallup Organization.

Zhang, M., Dorans, N., Li, C., & Rupp, A. (2017). Differential feature functioning in automated essay scoring. In H. Jiao & R. Lissitz (Eds.), *Test fairness in the new generation of large-scale assessment* (pp. 185–208). Charlotte, NC: Information Age Publishing.

Zheng, A. (2015). *Evaluating machine learning models: A beginner's guide to key concepts and pitfalls*. Sebastopol, CA: O'Reilly Media.