Contents

Machine-Scored Items

AIR has developed a large suite of machine-scored item types that provide fast scoring across a range of selected- and constructed-response item types. Most machine-scored item types are scored using explicit, test developer-defined rubrics that are contained in the item, which ensures fidelity to scoring rules across all test administrations. The rule-based rubrics allow test developers to construct test items that measure reason-based problem-solving, and validity of test score interpretation is increased because the rubrics explicitly define how students must respond to achieve each score point rather than providing guidelines for the kinds of responses associated with each score point.

AIR has six scoring engines that use explicit rubrics:

1. Multiple-choice scoring engine, which takes an option identifier as the key
2. Hot-text scoring engine, in which students select or rearrange sentences or phrases in a passage
3. Graphic-response scoring engine, which has an explicit test developer-created rubric that describes the properties of correct responses and relates scores assigned to those properties
4. Equation scoring engine, which evaluates the characteristics of student-entered equation responses against an explicit test developer-created rubric
5. Proposition-response scoring engine, which uses a pattern-matching algorithm to recognize test developer-created formal propositions in text using varying words, grammar, etc. (The explicit rubric defines concepts and relations among them and the relationship between the presence or absence of propositions and scores assigned.)
6. Simulation-interaction scoring engine, which evaluates a sequences of trials in a simulation item against an explicit test developer-supplied rubric

We propose to use our Automated Essay Scoring (AES) engine, which uses statistical algorithms to score student essays for the writing assessment in real time. The statistical rubrics used to develop the scoring models measure a broad set of features, some of which may be item specific and “learned” from a training set. During training, these features are related to human scores through a statistical model. The resulting equation predicts how a human would score a response with the measured features.

Autoscore is currently being used to accurately and quickly score student essay responses in Arizona, Ohio, and Utah. Statistical scoring models have already been developed to score most AIRCore writing assessments, so automated scoring of essay responses for New Hampshire’s statewide assessments can begin immediately. In addition, we note that there are additional AIRCore writing assessments for which scoring models have not yet been developed. Since reporting of assessment results for spring 2018 must necessarily await the outcome of standard setting workshops, the Department may also choose to administer writing tasks for which models have yet to be developed. Following the procedures described below, a sample of essay responses for training and cross validation will be identified, and for which optimally valid human scores will be obtained. From the responses, statistical scoring models will be developed allowing for immediate scoring and reporting of essay responses for those items in spring 2019. In the following paragraphs we describe how the AES scores student essays, and how we maintain the validity of all writing scores.

AES uses a statistical algorithm to score essay responses. The statistical scoring models also yield an indicator of score confidence based on (1) responses with unusual features and (2) responses scoring near rubric thresholds. For each model, a confidence threshold can be identified, and any scored response with a confidence value below the threshold will be automatically routed for verification scoring by a trained reader.

Scoring rubrics for all machine-scored items are highly secure, because they never leave our host servers. Although item scoring is nearly instantaneous, a modular, asynchronous scoring framework allows the

adaptive algorithm to proceed with item selection even while awaiting completion of scoring of submitted item responses, both machine scored and handscored. This ensures that students never need to wait for items to be scored to proceed through the test, even if demands on the network slow the scoring of some machine-scored items or item responses require handscoring following test administration.

AIR proposes 100% machine scoring of constructed-response items for all New Hampshire assessments. In conjunction with pre-equated item parameter estimates used for both item selection in the adaptive algorithm and immediate reporting of summative and interim assessments in the ORS and additional interim assessment reporting in AIRWays, machine scoring allows for immediate reporting of all test results. Upon submission of the test, the student's test record is passed to the Quality Monitor (QM), which performs a series of checks on the test record, including item and test scores. Records passing successfully through the QM (which are virtually all of them) are deposited to the database of record (DoR), where they are available to the Online Reporting System (ORS) and, for interim assessments, AIRWays.

As noted above, when assigned scores are near rubric thresholds, or when responses are dissimilar from those encountered in the training set, the confidence value assigned to the score becomes low. For accountability assessments, we propose to flag very low confidence scores automatically and route those responses for verification scoring by human raters. We propose to flag approximately 15% of all responses for verification reads.

As we describe in Section D1.6 Reporting, and Section D1.8 Reporting Portal, the ORS dynamically creates score reports from the contents of the DoR and the roster tracking system (RTS). As soon as an individual student's test record is passed to the DoR, his or her assessment results will be available to the reporting portals (the ORS and AIRWays). Aggregate-level reports will also be available, with class, school, and district reports being updated as tests are completed and become available for reporting.

Automated Scoring of Student-Generated Responses

Creating and Refining Machine Rubrics for Scoring Engines with Explicit Rubrics

Rubrics for each explicit rubric scoring engine are first developed by item writers and reviewers. We note that our rubrics support true rule-based reasoning, which is much more flexible and powerful than the simple token-matching approaches that are more common. True rule-based items allow more flexible scorings and admit rubrics that are more tightly aligned to the constructs being measured. These types of rubrics can be used for true constructed-response items. As with any constructed-response items, examinees may generate responses not anticipated by the test developers. Therefore, each item goes through a process similar to rangefinding for human-scored items.

AIR has developed a process called *rubric validation* that efficiently reviews scoring rubrics for true rule-based scoring. This process is supported by our REVISE software.

AIR's proven method for rubric validation is second-to-none in the industry. After items are field tested, AIR content specialists work with the Analysis Team to prepare for the rubric validation meetings. All student responses for non-selected response items are loaded into AIR's REVISE system, and AIR content specialists use the REVISE system to generate student samples based on a stratified random sample: one-third of the samples generated reflect students who performed well on the overall assessment but poorly on the item in question, one-third reflect students who were less proficient overall on the assessment but performed relatively well on the item in question, and one-third reflect students who performed as expected on the item in question. This process allows AIR to identify any potential scoring concerns prior to the rubric validation meeting, such as unanticipated (but accurate) responses, equivalent responses that were not originally considered, and responses that are getting credit but should not (based on the content and the item rubric).

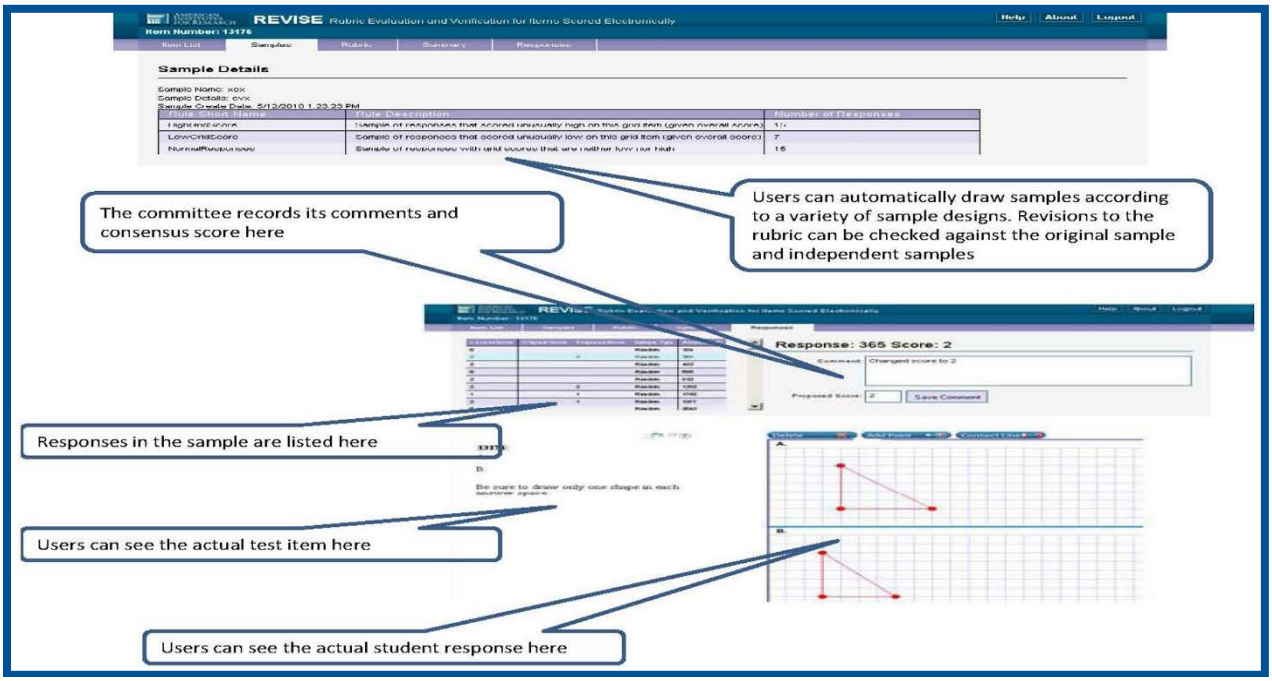
During the rubric validation meeting, AIR content specialists lead a training for participants to explain the rubric validation process, the importance of teacher participation and review of student responses, and how to review student responses in light of the current rubric. Participants are grouped by grade and content area, and AIR content specialists facilitate in each meeting room. Participants read the item/question, review the current rubric, and review the stratified random sample of student responses. If participants note responses that should be receiving credit but are not currently, they can recommend to Department that the rubrics be updated to include these additional correct responses. Similarly, if there are responses that participants feel should be excluded, they can be disallowed in the updated scoring. Finally, if a multi-point item is not performing well as a multi-point item, AIR may recommend to Department that the item be collapsed to a one-point item. (For example, if 48% of students previously earned 0 points, 48% of students earned 2 points, and only 4% of students earned 1 point, it could be argued that the item should truly be a 1-point item.)

Department content specialists will approve all changes to rubrics before the item is finally approved or rejected. AIR content specialists who are well-versed in item scoring and rubrics will apply the approved edits and generate all student responses whose scores have changed as a result of these rubric edits. These responses are reviewed to ensure that new responses deemed correct match the new rubric, and student responses that are no longer deemed correct are not receiving credit.

After the rubric validation meetings, any items rejected as part of the process (and approved by the Department) are rejected within ITS and are not moved forward. All items that are accepted at rubric validation are rescored with the updated rubrics by AIR's analysis team, and then AIR psychometric staff review the item statistics of the items (based on the updated rubrics and scoring) prior to data review.

Exhibit D1.5-1 highlights some of the features of the REVISE software.

Exhibit D1.5-1: Features of the Revise System



For the equation and proposition items, we bolster the rubric validation process with a validation study that compares the performance of the machine-scoring rubric with handscoring. To execute the validation study, a random sample of 500 cases from each item are handscored, and discrepancies from the machine score are reviewed and resolved. We will report a validity rate for each item.

Automated Essay Scoring Training

The engine employs a training set, a set of essays scored with optimally valid scores, which we obtain by having all responses double- scored by expert scorers and using an adjudication process for discrepant scores. The quality of the human-assigned scores is critical to the identification of a valid model and final performance of the scoring engine. The essay scoring engine is a statistical scoring engine. Measurable features of the essays are extracted and statistically analyzed. The measurable features include both syntactic features (e.g., misspellings, sentence length, grammar mistakes) and semantic features.

The semantic features reflect a summary of the words used in the documents. The summary is formed by first reducing words to their roots (e.g., "eating" becomes "eat") and creating a matrix of words (or n-grams, which are sequences of words) by responses. This, of course, is a very large sparse matrix. The matrix is factor analyzed to reduce its dimensionality, and the factors become the semantic predictors.

During training, the first step separates a random sample of approximately 500 cases to reserve for independent validation of the final model.

The second step removes responses flagged for deterministic condition codes. The computer, of course, is quite good at identifying responses that are too short to score, in which the prompt is copied, or where the student simply copies the same text over and over. These cases are flagged and removed from the training. The remaining cases, which still include cases to which humans assigned other condition codes, such as "off topic," are decomposed into the variables that comprise the predictors in the model. These predictors, or subsets of them, are included in ordered probit models predicting (1) whether a condition code would be applied, and (2) the score that would be applied. Each prediction comes with a confidence index. An algorithm takes this information into account in determining whether to assign a condition code or a score.

When our psychometricians are satisfied with the models, they are run against the validation sample to estimate the agreement between the scoring engine and the fully resolved human score.

The essay scoring system generates several confidence indices, which are then used to determine whether the paper should receive a score or a condition code. The mean and standard deviation of the confidence index are computed for each dimension, and the lowest value is selected for flagging responses with one or more low confidence scores. Any dimension score with a confidence index below this threshold is flagged for verification by a human rater.

We propose to route papers with the lowest confidence scores to human scorers. In the past this has amounted to approximately 20% of the papers with the expected turnaround being seven business days. This would mean that 80% of students get immediate scores in AIR's reporting systems and the other 20% get their scores within seven business days.