ETS®

Listening. Learning. Leading.®

R & D Connections

No. 13 • February 2010

Constructed-Response Scoring — Doing It Right

by Catherine A. McClellan

How do we standardize the scoring of responses to performance assessment tasks — which assessment people often refer to as *constructed-response* (CR) *items*[1] — so that the scores are reliable and so that they have the same valid meaning for all test takers?

We expect scores from standardized tests to be comparable over time and over different administrations and forms of the test. For assessments composed of multiple-choice items, there are a number of techniques to accomplish this, such as fixed timing, machine-scored answer sheets, equating different forms, and scores reported on a scale rather than number or percent correct. For such tests, it should not matter to the examinee which form of a test he or she takes on which occasion.

But what about tests composed, in part or in full, of questions requiring examinees to write essay responses, show the steps in solving a math problem, create pieces of art, perform dances, or record spoken language — in short, anything that requires an examinee to construct a response (hence the name ***constructed response***), rather than select one provided on the test?

These so-called CR items are usually scored by people, and standardizing people's judgments and actions is not a simple matter. But in testing, lack of standardization can lead to all sorts of bad outcomes.

Critical Elements of Standardized Testing

Two essential properties for tests are validity and reliability. Reliability is whether the measure gives a consistent picture of performance for each examinee (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999; Haertel, 2006). Validity is whether the decisions we make using the test scores lead to the intended outcomes (AERA, APA, & NCME, 1999; Kane, 2006).

What is involved?

These are some important components in performance testing:

- *Constructed-response items (often known as performance tasks)* — A task that requires test takers to construct answers rather than select from predetermined multiple-choice options; examples include essays, works of art or speeches.
- *Rubric* — The set of scoring standards that describes the criteria for each score level.

> *"If we aren't measuring something consistently, we cannot use that measurement to make an appropriate decision about the examinee."*

[1] The terms "prompt" and "item" are used in this article to describe tasks that elicit a constructed response from a test taker; often, the prompt or item does not come in the form of a question.

__Editor's note:__ Catherine A. McClellan is the director of human constructed-response scoring in the Research Applications & Development area of ETS's Research & Development Division.

Who is involved?

Other than the test takers, these are some important people in the process of constructed-response scoring:

- *Raters* — People hired to score constructed-response tests.
- *Scoring leader* — An experienced rater who has shown consistently strong scoring performance and who has the interpersonal qualities of a good mentor.

The two properties are related: If we aren't measuring something consistently, we cannot use that measurement to make an appropriate decision about the examinee. A test can be reliable without being valid: For example, we could measure examinees' height very reliably, but this likely is not a valid indicator of the same examinees' writing proficiency.

The opposite, on the other hand, is not true: An unreliable test cannot be valid.

Now imagine the human beings — who we refer to as ***raters*** — hired to score standardized writing tests. Obviously, they have to know something about good writing. This is a necessary, but not a sufficient, condition for their scores to be *valid*. In addition to this expertise, these raters also have to be able to give similar scores to similar examinee responses, and their scores have to be similar to those of other raters, as well.

Even if we select people with knowledge of writing to rate essay tests — perhaps they have college degrees in English — how would we know if the scores they assign to the essays are reliable? Is one rater's judgment of writing quality different from another rater's? How do we know? How do we assure that the scores assigned to the essays are comparable to each other? If they aren't, then the scores on the test are no longer comparable to each other either. Two examinees with precisely the same responses could receive different scores on a test due solely to what different raters think the response deserves. If decisions about the examinees are made based on flawed scores such as this, the decisions will be flawed as well.

To have a truly *standardized* test, it must not matter which rater scores an examinee's responses any more than it matters which test form is taken on which occasion.

Controlling Rating Quality

Without well-written prompts that elicit a broad variety of responses and a scoring rubric that clearly defines each distinct score level, high-quality scoring will not occur. Even given those conditions, some methods of ensuring standardization of raters are obvious; some less so.

Credentials are an obvious criterion: Raters must have knowledge of the content area in which they are scoring responses. How to assure this may be less obvious, although in practice, this typically is established through educational or professional credentials.

> *"...content knowledge alone is not sufficient — a rater must be trained in the specific procedures needed to score the responses to the particular test or item he or she will be scoring."*

But content knowledge alone is not sufficient — a rater must be trained in the specific procedures needed to score the responses to the particular test or item he or she will be scoring. CR items on standardized tests typically have a fixed set of score levels to which responses will be assigned, and detailed descriptions of the performance that qualifies a response for each level. This information is described in a document called a scoring rubric. See Lane & Stone (2006) for a summary description of CR scoring procedures.

In order to have consistent and reliable CR scoring, each rater must understand and apply the scoring rubric to the examinee responses in the same way every time.

Raters receive extensive training on the scoring rubric for the item to be scored. This training may come from an instructor or from a self-paced online tutorial.

Described next is a common set of steps leading up to live scoring. As part of the training, raters receive sample responses with the score level pre-assigned. Training focuses on the reasons each response received the score level it did. Next, raters may receive responses to discuss and assign scores to, either individually or as a group. Once the raters are comfortable with the scoring rubric, each rater individually assigns scores to a set of responses that have previously received a score from content experts. After everyone completes this scoring, the training leader will poll the raters for their scores on each response and discuss the pre-assigned score and the rationale for assigning it. If the rater training is online, the discussion is managed through the use of extensive commentaries and descriptions instead of orally.

As a culminating exercise to verify that each rater has understood the scoring rubric and can apply it effectively, a set of responses is given to test the rater's skills — so raters themselves are tested before they are allowed to score tests. Each rater must score these calibration responses individually. Raters who do not reach a specified level of agreement are retrained and given one more chance to successfully calibrate; if a rater fails twice, he or she may not score operationally for that shift. Raters who satisfy the criterion for ***calibration*** may move on to operational scoring. Training is done at the beginning of the scoring session for an item, and may take many days to finish; a shorter version of the training may be repeated before each scoring shift. To assure ongoing quality in scoring, raters have to recalibrate at the beginning of every shift of scoring.

"Even after rigorous training and calibration, rater scoring performance is not taken for granted."

Even after rigorous training and calibration, rater scoring performance is not taken for granted. Assessment developers, statisticians and, especially, scoring leaders evaluate the quality of operational scoring as it is occurring through a number of statistical and qualitative approaches to assure any problems are caught and corrected as soon as possible (NCES, 2008). The scoring leaders play a crucial role in promoting reliability of the scores and creating conditions that allow score users — such as college admissions offices or teacher licensure bureaus — to make valid inferences about the meaning of those scores.

Monitoring scores

Raters work in small teams, typically of 6 – 10 people, under the direction of a scoring leader. This leader is an experienced rater who has shown consistently strong scoring performance, in addition to having the interpersonal qualities to be a strong mentor. Scoring leaders monitor team performance throughout the scoring shift. Supplementing the scoring leader's guidance, several techniques are used by other specialists to evaluate the quality of the scoring (NCES, 2008).

One key approach to score monitoring is called ***backscoring***. In this approach, the scoring leader evaluates responses that raters have already scored and verifies that they did not assign the scores in error — "back" scoring because the leader is following along behind the raters and reviewing their work. The scoring leader backscores throughout the scoring shift for all raters on the team, randomly sampling

Approaches to Monitoring Scores

These are some common ways of controlling rating quality:

- *Backscoring* — Experienced scoring leaders review responses that raters have already scored and verify that they have not assigned the scores in error.
- *Using validity responses or monitor responses* — Content experts include clear and unambiguous examples of a particular score level among the actual responses. Raters score responses without knowing which ones are actual responses and which ones are control responses.
- *Double scoring* — Two raters score the same response, allowing scoring leaders to see whether raters in the same scoring session are generally applying the scoring rubric in the same way.
- *Trend scoring* — Test analysts check to see whether raters are applying the scoring rubric the same way over time, from one test administration to the next.

responses. Discrepancies result in a conference between the scoring leader and the rater to correct the rater's misunderstanding and misapplication of the scoring rubric, so that the rater will score correctly going forward. Incorrectly assigned scores are corrected as they are found in backscoring. If further backscoring of a particular rater indicates a persistent problem, that rater can be removed from operational scoring; all scores assigned by that rater can be cancelled and those responses assigned to other raters for a new score. These are serious actions, costly in both time and resources, and are taken only after consultation by the scoring leader with the assessment team (usually the test developer, a program representative and a psychometrician).

Another tool used to monitor the quality of CR scoring is ***validity responses***, which are also known as ***monitor responses***. Content experts — generally assessment developers and scoring leaders — select these exemplars as clear and unambiguous examples of a particular score level. Validity scores are verified independently by a minimum of two experts, so that to the extent possible, the score on a validity response is a correct score. For each item being scored, a set of validity responses is chosen so that there are examples of all score levels in the scoring rubric and of all types of common responses to the item. These responses are seeded into the operational scoring, and raters assign scores to the validity responses as part of the regular scoring. If at all possible, the validity responses appear to be just one more response to the raters, with nothing that indicates the response is a validity one. This is to ensure that the evaluation is of the normal scoring behavior of the rater, unaltered by the knowledge that he or she is being "tested."

The scoring leader examines the scores that raters assign to the validity responses. Any discrepancies between the correct scores and the rater-assigned scores also trigger a conference between the scoring leader and rater to correct the rater's misunderstanding and misapplication of the scoring rubric, so that the rater will score correctly going forward. Incorrect scoring of validity responses triggers increased backscoring from the scoring leader for that rater to ensure that scoring quality standards are maintained. The scoring leader recommends or provides additional training as necessary, and can dismiss the rater from operational scoring if necessary.

The number of raters who score each response varies by testing program. If a response is scored by two raters, it is referred to as a ***double-scored response***. Some percentage of responses in many testing programs is double scored as a quality measure; in some high-stakes testing programs, all responses are double scored.

Double-scored responses all receive a statistic, calculated as a quality measure, called ***interrater*** agreement. This statistic refers to the frequency with which two raters assign the same score to the same response; the double scoring provides the data necessary to calculate the value of this statistic.

Trend scoring is a way of checking that raters are evaluating responses consistently from one test administration to the next. Basically, this score monitoring method involves taking a set of examinee responses from a previous administration and having the raters for the current administration rescore them. This method is particularly useful when the groups of raters change from administration to administration.

Why Monitor?

Backscoring, validity scoring, double scoring and trend scoring differ in some important ways — most critically, in the type of information provided and the conclusions each one supports about the quality of scoring. Each approach checks a slightly different assumption about the quality of scoring.

- *Backscoring* is the most individualized and the most qualitative monitoring step. The object of the feedback is the individual rater being backscored by the scoring leader. The response is immediate and very specific to the actions of one specific rater and the reasons a particular score was assigned to a response. Discussion and clarification of how the scoring rubric should be applied improve the individual rater's scoring.

- *Validity scoring* allows the current scoring to be held up to the standard of truth of perfect scoring (to the extent that is feasible). Validity responses are selected based on the belief that experts can assign an unambiguous correct score to each response in the validity set. A validity response is an exemplar of a score level, and a well-trained rater who understands the scoring rubric and how to use it should not have difficulty in recognizing the correct score to assign. Validity scoring provides quality indicators at both the individual and the rater group level. Deviations from the correct score are an indication of rater misunderstanding, and indicate a need for prompt intervention by the scoring leader with the rater. A review of the summary performance of a group of raters on the set of validity papers scored during a scoring session provides valuable feedback on overall operational scoring quality.

- *Double scoring* gets at something quite different from validity scoring. This approach allows verification of whether raters within a scoring session are agreeing with each other on scores. While this may sound a lot like validity scoring, there is an important difference: We do not know if *either* rater in double scoring is correct. Raters can agree and both be wrong in the same way. Double scoring tells us that the raters are consistently applying the scoring rubric in the same way — not if they are applying it correctly. Double scoring also provides quality indicators at both the *individual and the rater group levels*. Double scoring does have an advantage in that the validity set of responses is usually fairly small and is not representative; validity responses don't represent borderline cases, as they are chosen to be clear exemplars of each score level. The double-scored set should be randomly selected from the operational responses, and so will cover a broader range of response types and borderline cases. Evaluating whether or not raters agree on how to score the difficult judgment cases is indispensable in determining how effective the rater training has been in refining the raters' understanding of how to apply the scoring rubric.

- *Trend scoring* checks on yet another aspect of scoring. In this case, it is consistency of the scoring rubric application over time that is validated. Trend scoring does not check raters against a standard of correctness, nor does it check them against the current scoring — rather, trend scoring considers history. If there are deviations from the historically assigned scores in the current scores assigned, the reason should be investigated and, if necessary, the problem corrected. Trend scoring provides information about the performance of the current group of raters compared to a past group of raters. The score assigned to the same response can ***drift*** over time due to alterations in training materials, shifts in the emphases given to particular parts of training, or changes in the raters scoring the tests. These types of shifts in scores assigned to responses must be corrected through adjustments in the training provided to the raters. If the scores assigned to responses to an item change over time for a legitimate reason (and this can happen), a decision as to whether or not the item should be included as part of the measure over time must be made by assessment content experts. Content that shifts over time occurs most frequently in the sciences, where professional consensus about facts may evolve. For example, new discoveries or revised definitions may change previously incorrect responses to correct — or *vice versa* — on topics such as what diseases have a hereditary component or whether Pluto is a planet.

Costs and Benefits

Correctly managing the scoring of CR items requires a lot of attention and good controls. Since human beings, rather than scanners, score the responses — and human beings get paid for this work (while scanners don’t) — CR items are very expensive indeed to have on a test. If a test has CR items on it, it is a safe bet that the amount of time it takes to report the scores is driven largely by how long it takes to score the CRs. Expensive, time-consuming and difficult to do well — why does anyone use CR items, anyway?

“Expensive, time-consuming, difficult to do well — why use constructed-response items, anyway? The short answer: Validity.”

The short answer is validity. Validity is, in many ways, the ultimate characteristic of assessment. To quote the ***Standards for Educational and Psychological Measurement*** (AERA, APA, & NCME, 1999), “Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests.” CR items are believed to measure some educationally important skills and types of knowledge — often complex skills, direct performances or explication of reasoning — which are not well measured with multiple-choice questions (see Livingston, 2009).

“CR items are believed to measure some educationally important skills and types of knowledge which are not well measured with multiple-choice questions.”

There have been major advances in using computer software to score certain types of constructed responses, such as writing and speech — an area referred to as ***automated scoring***. Use of these systems may reduce the turnaround time for scoring and is less costly than paying human raters for the same work. The automated scoring systems in operational use are very accurate in matching the scores humans assign to the same responses, although the processes by which the machine and the human arrive at that score may differ. Automated scoring models are most often created by “training” the system on a set of double-human-scored responses, and for high-stakes assessment, the automated scoring systems customarily perform second scoring after a first human score is applied, so human scoring has not been eliminated from CR scoring just yet. A more complete discussion of automated scoring is beyond the scope of this work, but can be found in Shermis, Burstein, Higgins, & Zechner (in press).

> *"CR items are popular choices in standardized educational assessment, so it is important to assure that the human scoring aspect of the tests, just as all others, is as standardized as possible."*

Summary

The perceived benefits of constructed-response versus multiple-choice items (Livingston, 2009) must be weighed carefully when constructing a test. Despite the associated limitations in costs, time and complexity imposed in scoring them, CR items remain popular choices in standardized educational assessment. Given that, it is important to assure that the human scoring aspect of the tests, just as all others, is as standardized as possible. Without careful item and rubric construction, thorough rater training and skills evaluation, and vigilant monitoring of scoring quality, the use of CR items will have a detrimental impact on test reliability and score validity. When used selectively and scored with rigor, CR items provide valid information and insight into students' achievements on complex performance skills.

References

American Educational Research Association, American Psychological Association & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 65-110). Westport, CT: American Council on Education/Praeger.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17-64). Westport, CT: American Council on Education/Praeger.

Lane, S., & Stone, C. A. (2006). Performance assessment. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 387-432). Westport, CT: American Council on Education/Praeger.

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

National Center for Education Statistics. (2008). *NAEP technical documentation: Scoring monitoring.* Retrieved July 10, 2009, from http://nces.ed. gov/nationsreportcard/tdw/scoring/scoring.asp

Shermis, M., Burstein, J., Higgins, D., & Zechner, K. (in press). Automated essay scoring: Writing assessment and instruction. In *International encyclopedia of education* (3rd ed.). New York: Elsevier.

R&D Connections is published by

ETS Research & Development
Educational Testing Service
Rosedale Road, 19-T
Princeton, NJ 08541-0001
e-mail: **RDWeb@ets.org**

Editor: Jeff Johnson

Visit ETS Research & Development
on the Web at
www.ets.org/research