Listening. Learning. Leading.®

R&D Connections

No. 11 • September 2009

Constructed-Response Test Questions: Why We Use Them; How We Score Them

By Samuel A. Livingston

To many people, standardized testing means multiple-choice testing. However, some tests contain questions that require the test taker to produce the answer, rather than simply choosing it from a list. The required response can be as simple as the writing of a single word or as complex as the design of a laboratory experiment to test a scientific hypothesis.

These types of test questions, taken together, are referred to as *constructed-response* questions.

What kind of information can we get from constructed-response questions that we cannot get from multiple-choice questions?

How do we translate the test takers' responses into numerical scores?

What are the limitations of constructed-response questions, and how are recent technological developments helping to overcome some of those limitations?

Beyond Multiple-Choice

The multiple-choice question format has come to dominate large-scale testing, and there are good reasons for its dominance. A test taker can answer a large number of multiple-choice questions in a limited amount of testing time. The large number of questions makes it possible to test a broad range of content and provides a good sample of the test taker's knowledge, reducing the

Editor's note: Samuel A. Livingston is a senior psychometrician in the Statistical Analysis and Psychometric Research area of ETS's Research & Development division.

Examples

Constructed-response questions are a way of measuring complex skills. These are examples of tasks test takers might encounter:

- *Literature* — Writing an essay comparing and contrasting two poems, stories, or plays
- *Mathematics* — Writing a mathematical equation to solve a problem presented in words and diagrams
- *Biology* — Describing how a biological process occurs in a plant and explaining how it enhances the plant's ability to survive or to reproduce
- *Music* — Listening to a melody or a chord progression and writing it correctly in musical notation
- *History* — Writing an essay comparing two instances of a social or political process that occurred at different times in different regions of the world

effect of "the luck of the draw" (in the selection of questions) on the test taker's score. The responses can be scored by machine, making the scoring process fast and inexpensive, with no room for differences of opinion.

No wonder that large-scale testing organizations have come to depend heavily on multiple-choice questions. Why would anyone ever want to use anything else?

One reason is that many skills that schools teach are too complex to be measured effectively with multiple-choice questions.

A multiple-choice test for history students can test their factual knowledge. It can also determine whether they can discriminate between correct and incorrect statements of the relationships between facts — but it cannot determine whether the students can write a well-reasoned essay on a historical question.

The tasks that constructed-response questions require of the test taker are as varied as the skills to be measured.

A multiple-choice test for mathematics students can determine whether they can solve many kinds of problems — but it cannot determine whether they can construct a mathematical proof.

A multiple-choice test of writing ability can determine whether the test takers can discriminate between well written and badly written versions of a sentence — but it cannot determine whether they can organize their own thoughts into a logically structured communication in clear and appropriate language.

Another reason for using constructed-response questions is that a test taker who can choose the correct answer from a list may not be able to provide the answer without seeing it presented. Is the difference educationally important? Sometimes it is.

Students who cannot remember the correct procedure for conducting a science laboratory experiment may recognize the correct next step, or the correct sequence of steps, when they see it.

Students who cannot explain the logical flaw in a persuasive message may find it easy to identify the flaw when it is presented as one of four or five possibilities.

Students who cannot state the general scientific principle illustrated by a specific process in nature may have no trouble recognizing that principle when they see it stated along with three or four others.

Making the multiple-choice questions more difficult by making the wrong answers more like the correct answer does not overcome this limitation. Instead, it can cause some test takers who know the correct answer (without seeing it presented) to miss the question by choosing one of the nearly-correct answers instead.

Measuring Complex Skills

The tasks that constructed-response questions require of the test taker are as varied as the skills to be measured — a wide variety, even if the skills are limited to academic skills that can be demonstrated with a pencil and paper. (Tasks that require responses that cannot be made in pencil-and-paper format are typically described as "performance assessments," rather than "constructed-response questions.")

In literature, the test taker may be asked to write an essay comparing and contrasting two poems, stories, or plays.

In mathematics, the test taker may be asked to write a mathematical equation to solve a problem presented in words and diagrams.

In biology, the test taker may be asked to describe the way a particular biological process occurs in a type of plant and explain how it enhances the plant's ability to survive or to reproduce.

In music, the test taker may be asked to listen to a melody or a chord progression and write it correctly in musical notation.

In history, the test taker may be asked to write an essay comparing two instances of a social or political process that occurred at different times in different regions of the world.

The additional capabilities of constructed-response test questions come at a price.

The additional capabilities of constructed-response test questions for measuring complex skills come at a price.

These kinds of constructed-response questions take longer to answer than multiple-choice questions, so that a test taker cannot answer as many of them in a given amount of time. Consequently, an individual test taker's performance will tend to vary more from one set of questions to another.

The responses tend to be time-consuming to score, increasing the cost of testing and the time required to compute and report the scores.

The scoring process often requires judgment, so that different scorers can possibly award different scores to the same response.

Multiple-Choice as a Substitute for Constructed-Response

Some people in the field of educational testing have claimed that multiple-choice and constructed-response questions provide essentially the same information. Therefore, they argue, multiple-choice questions can be used as a substitute for constructed-response questions (Lukhele, Thissen & Wainer, 1994).

These claims are based on research studies showing a high level of agreement between scores on multiple-choice and constructed-response questions (e.g., Godschalk, Swineford, & Coffman, 1966). However, those research studies generally have compared the multiple-choice and constructed-response scores of a single group of test takers who were tested once with both types of questions. The high level of overall agreement can mask important differences between groups of test takers.

For example, research studies have shown that male/female differences on constructed-response questions often do not parallel the male/female differences on multiple-choice questions in the same subject (Mazzeo, Schmitt, & Bleistein, 1992; Breland, Danos, Kahn, Kubota, & Bonner, 1994; Livingston & Rupp, 2004). Typically, when women and men perform equally well on the multiple-choice questions, the women outperform the men on the constructed-response questions. When women and men perform equally well on the constructed-response questions, the men outperform the women on the multiple-choice questions. These differences occur even though the multiple-choice scores and the constructed-response scores tend to agree strongly within each group.

Scoring Guides

Both main approaches to scoring constructed-response items — analytic scoring and holistic scoring — use sets of guidelines called *rubrics*.

Analytic scoring rubrics:

- List specific features of the response
- Tell the scorer how many points to award (or subtract) for each specific feature
- Tend to produce scoring that is highly consistent from one scorer to another

Holistic scoring rubrics:

- Contain statements describing the characteristics of a typical response at each score level
- Require the scorer to assign a single numerical score to the whole response
- Are used in conjunction with *exemplars* — actual responses selected as examples for each possible score
- Can be used when it is not possible to describe the quality of a response in terms of specific features that are either present or absent

A high level of agreement between two measures does not mean that the two measures will change in the same way over time. In a group of school children, height and physical strength tend to agree strongly; the tallest students tend to be the strongest. However, a three-month intensive physical training program will increase the students' strength, without causing any change in their height. Measuring the students' height before and after the training program will not show the increase in their strength.

Similarly, in academic subjects, there is usually a strong tendency for the students who are stronger in the skills measured by multiple-choice questions to be stronger in the skills measured by constructed-response questions. But if all the students improve in the skills tested by the constructed-response questions, their performance on the multiple-choice questions may not reflect that improvement.

This limitation of multiple-choice test questions has educational consequences. When multiple-choice tests are used as the basis for important decisions about the effectiveness of schools, teachers have a strong incentive to emphasize the skills and knowledge tested by the questions on those tests. With a limited amount of class time available, they have to give a lower priority to the kinds of skills that would be tested by constructed-response questions.

Scoring the Responses

There are two basic approaches to the scoring of constructed-response test questions. These approaches are called *analytic scoring* and *holistic scoring*. In both cases, the scoring is based on a set of guidelines called a *rubric*. The rubric tells the scorer what features of the response to focus on and how to decide how many points to award to the response.

An analytic scoring rubric lists specific features of the response and specifies the number of points to award for each feature. On a question in applied mathematics, the scorer may award one point for identifying the relevant variables, one point for writing an

equation that will solve the problem, and one point for solving the equation correctly.

On a science question, the scorer may award two points for providing a correct explanation of a phenomenon, one point for correctly stating the general principle that it illustrates, and one point for providing another valid example of that principle in action.

The process of holistic scoring is very different. The scorer reads the response and makes a single judgment of the quality of the response by assigning a numerical score.

A holistic scoring rubric usually contains statements describing the characteristics of a typical response at each score level. However, to define the score levels in practical terms that the scorers can apply requires *exemplars* — actual responses written by test takers, selected as examples of a 5-point response, a 4-point response, etc.

The exemplars also include borderline cases — for example, a response that just barely qualifies for a score of 5, or a response that narrowly misses earning a score of 5.

Analytic scoring tends to be more consistent from one scorer to another than holistic scoring; the same response, scored by two different scorers, is more likely to receive the same score from both scorers. Analytic scoring works well when:

- the question designer can explicitly specify the features of the response for which test takers should receive points (or, in some cases, lose points), and
- the important features of the response can be evaluated separately; the quality of the response does not depend on interactions among those features.

In a purely analytic scoring system, the scoring criteria can be expressed as a set of yes-or-no questions. (Did the student correctly identify the scientific principle? Did the student provide another valid example?)

Some analytic scoring systems bend slightly in a holistic direction, allowing the scorer to award partial credit for some features of the response — for example, 2 points for a fully correct explanation; 1 point for a partly correct explanation.

On some kinds of constructed-response questions (e.g., questions intended to test writing ability), it is not possible to describe the quality of a response in terms of specific features that are either present or absent. Responses to these questions are scored holistically.

One of the greatest problems in constructed-response testing is the time and expense involved in scoring. The process often includes elaborate systems for monitoring the consistency and accuracy of the scores.

Automated Scoring

One of the greatest problems in constructed-response testing is the time and expense involved in scoring. The scoring process requires substantial amounts of time from highly trained scorers and often includes elaborate systems for monitoring the consistency and accuracy of the scores.

In recent years, researchers have made a great deal of progress in using computers to score the responses. Automated scoring offers the possibility of greatly decreasing the time and cost of the scoring process, making it

Automated Scoring Engines

Automated scoring offers the possibility of decreasing the time and cost of scoring constructed-response questions. ETS has developed four scoring engines:

- The *e-rater*® engine, which scores the quality of writing in essays by predicting the score that a human scorer would assign
- The *c-rater*™ engine, which scores the content of written responses by scanning for statements that have been specified as correct answers or for other statements having the same meaning
- The *m-rater* engine, which scores the correctness of algebraic expressions, lines or curves on a graph, or geometric figures created by the test taker
- The *SpeechRater*SM engine, which scores the responses to a variety of speaking tasks that indicate the test taker's ability to communicate effectively in English

practical to use constructed-response questions in testing situations where human scoring would be impractical or prohibitively expensive.

Four *scoring engines* — computer programs for automated scoring — have been developed at ETS in the past few years. They are called:

- *e-rater*® (for "essay rater"),
- *c-rater*™ (for "content rater"),
- *m-rater* (for "math rater"),
- and *SpeechRater*SM.

The first of these to be developed was the *e-rater* scoring engine (Attali & Burstein, 2005). Its task is to assign to each test taker's essay a score that indicates the quality of the writing — the same score that an expert human scorer would assign.

The *e-rater* engine cannot actually read and understand the essay. What it can do is to record the linguistic features of the writing in the essay and use them to predict the score that an expert human scorer would assign. To prepare the *e-rater* engine for this task, the operators feed it a sample of essays representing a wide range of quality, along with the scores assigned to those essays by expert human scorers. The *e-rater* engine succeeds remarkably well at its task, producing scores that agree with the scores assigned by human scorers as closely as the scores assigned by different human scorers agree with each other.

The *c-rater* engine takes a very different approach (Leacock & Chodorow, 2003). Its task is to evaluate the content of the response, not the quality of the writing. The *c-rater* engine is designed for use with analytically scored short-answer questions. To score the responses to a particular question with the *c-rater* engine, it is necessary to enter into the computer the statements for which points will be awarded (or subtracted). The *c-rater* engine then scans the response for those statements — or for alternative statements having the same meaning. When it finds them, it awards the appropriate number of points.

The *m-rater* engine is a scoring engine for responses that consist of an algebraic expression (e.g., a formula), a plotted line or curve on a graph, or a geometric figure. To score an algebraic expression, the *m-rater* engine determines whether the formula written by the test taker is algebraically equivalent to the correct answer. The *m-rater* engine scores a straight-line graph by transforming the line into an algebraic expression; it scores a curved-line graph by testing the curve for correctness at several points.

SpeechRater, the newest of ETS's scoring engines, is still under development. The *SpeechRater* engine is intended for scoring the responses to a variety of speaking tasks that indicate the test taker's ability to communicate effectively in English (Zechner & Xi, 2008). Like the *e-rater* engine, it records linguistic features of the test taker's response and uses them to predict the scores that would be assigned by expert human scorers. And like the *e-rater* engine, it produces scores that tend to agree with the scores assigned by expert human scorers (though not as closely as the scores produced by the *e-rater* engine).

Summary

Multiple-choice test questions have many practical advantages, but they cannot measure some educationally important skills and types of knowledge. Some skills are too complex to be measured effectively with multiple-choice questions. Other skills and types of knowledge cannot be measured if the test-taker is shown a list that includes the correct answer.

Constructed-response test questions can measure many of these skills. However, constructed-response questions generally take longer to administer than multiple-choice questions. They are also much more expensive to score, and the scoring process often leaves room for differences of opinion.

When scores on multiple-choice and constructed-response questions agree closely, the multiple-choice questions may be an adequate substitute for the constructed-response questions.

However, the relationship between the multiple-choice and constructed-response scores can be different in two groups of test takers, such as men and women. Also, an improvement in the skills measured by the constructed-response questions may not be reflected in the scores on the multiple-choice questions.

The responses to constructed-response questions can be scored either analytically or holistically. Analytic scoring tends to produce scores that are more consistent from one scorer to another, but some kinds of constructed-response questions require holistic scoring.

A recent development in constructed-response testing is automated scoring — the scoring of the responses by computer. Automated scoring has the potential to make constructed-response test questions practical for use in situations where scoring by human scorers is not a practical possibility.

References

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® V.2. *Journal of Technology, Learning, and Assessment, 4*(3). Available from http://escholarship.bc.edu/jtla/

Breland, H. M., Danos, D. O., Kahn, H. D., Kubota, M. Y., & Bonner, M. W. (1994). Performance versus objective testing and gender: An exploratory study of an Advanced Placement history examination. *Journal of Educational Measurement*, *31*, 275-293.

Godschalk, F. I., Swineford, F., & Coffman, W. E. (1966). *The measurement of writing ability.* New York: College Entrance Examination Board.

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

Livingston, S. A., & Rupp, S. L. (2004). *Performance of men and women on multiple-choice and constructed-response tests for beginning teachers* (ETS Research Report No. RR-04-48). Princeton, NJ: Educational Testing Service.

Lukhele, R., Thissen, D., & Wainer, H. (1994). On the relative value of multiple-choice, constructed-response, and examinee-selected items on two achievement tests. *Journal of Educational Measurement*, *31*, 234-250.

Mazzeo, J., Schmitt, A., & Bleistein, C. A. (1992). *Sex-related performance differences on constructed-response and multiple-choice sections of Advanced Placement examinations* (College Board Research Rep. No. 92-7; ETS Research Report No. RR-93-05). New York: College Entrance Examination Board.

Zechner, K., & Xi, X. (2008). Towards automatic scoring of a test of spoken language with heterogeneous task types. In *Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 98-106). Columbus, OH: Association for Computational Linguistics.

R&D Connections is published by:

ETS Research & Development
Educational Testing Service
Rosedale Road, 19-T
Princeton, NJ 08541-0001

Send comments about this publication to the above address or via the Web at: **http://www.ets.org/research/contact.html**