Listening. Learning. Leading.®

R & D Connections

No. 21 • March 2013

Contrasting Automated and Human Scoring of Essays

By Mo Zhang[1]

Key Concepts

Constructed-response item: A test question that requires the test takers to supply the answer, instead of choosing it from a list of possibilities.

Common Core State Standards (CCSS): A set of curricular goals in English language arts and mathematics adopted by most states for students in grades K–12.

Large-scale testing programs: Assessments taken by a large volume of test takers, such as ETS's *TOEFL iBT*® test and Pearson's PTE.

High-stakes decision: A judgment, based in part on test results, that has significant consequences for an individual, a group, or an institution, such as college admission, graduation, and school sanctions.

Essay scoring has traditionally relied on human raters, who understand both the content and the quality of writing. However, the increasing use of constructed-response items, and the large number of students that will be exposed to such items in assessments based on the Common Core State Standards (CCSS), raise questions about the viability of relying on human scoring alone. This scoring method is expensive, requires extensive logistical efforts, and depends on less-than-perfect human judgment. Testing programs are therefore tapping into the power of computers to score constructed-response items efficiently. The interest in automated scoring of essays is not new and has recently received additional attention from two federally supported consortia, PARCC and Smarter Balanced, which intend to incorporate automated scoring into their common core state assessments planned for 2014.

Nonetheless, human labor cannot simply be replaced with machines, since human scoring and automated scoring have different strengths and limitations. In this essay, the two scoring methods are compared from measurement and logistical perspectives. Conclusions are drawn from research literature, including ETS research, to summarize the current state of automated essay scoring technology.

The published research has few in-depth comparisons of the advantages and limitations of automated and human scoring. There are also debates in academia, the media, and among the general public concerning the use of automated scoring of essays in standardized tests and in electronic learning environments used in and outside of classrooms. It is important for test developers, policymakers, and educators to have sufficient knowledge about the strengths and weaknesses of each scoring method in order to prevent misuse in a testing program. The purpose of this essay is to contrast significant characteristics of the two scoring methods, elucidate their differences, and discuss their practical implications for testing programs.

Human Scoring

Many large-scale testing programs in the United States include at least one essay-writing item. Examples include the GMAT® test administered by the Graduate Management Admission Council®, the *GRE*® revised General Test administered by ETS, as well as the Pearson® Test of English (PTE). The written responses to such items

[1] Editor's note: *Mo Zhang is an associate research scientist in ETS's Research & Development division.*

"Human labor cannot simply be replaced with machines since human scoring and automated scoring have different strengths and limitations."

are far more complex than responses to multiple-choice items, and are traditionally scored by human judges. Human raters typically gauge an essay's quality aided by a scoring rubric that identifies the characteristics an essay must have to merit a certain score level. Some of the strengths of scoring by human graders are that they can (a) cognitively process the information given in a text, (b) connect it with their prior knowledge, and (c) based on their understanding of the content, make a judgment on the quality of the text. Trained human raters are able to recognize and appreciate a writer's creativity and style (e.g., artistic, ironic, rhetorical), as well as evaluate the relevance of an essay's content to the prompt. A human rater can also judge an examinee's critical thinking skills, including the quality of the argumentation and the factual correctness of the claims made in the essay.

For all its strengths, human scoring has its limitations. To begin with, qualified human raters must be recruited. Next, they must be instructed in how to use the scoring rubric and their rating competencies must be certified prior to engaging in operational grading. Finally, they must be closely monitored (and retrained if necessary) to ensure the quality and consistency of their ratings. (See Baldwin, Fowles, & Livingston, 2005, for ETS's policies on performance assessment scoring.) In 2012, more than 655,000 test takers worldwide took the GRE revised General Test (ETS, 2013), with each test taker responding to two essay prompts, producing a total of more than 1.3 million responses. Obviously, involving humans in grading such high volumes, especially in large-scale assessments like the GRE test, can be labor intensive, time consuming, and expensive.

Humans can also make mistakes due to cognitive limitations that can be difficult or even impossible to quantify, which in turn can add systematic biases to the final scores (Bejar, 2011).

Table 1 exemplifies sources of human error known from the research literature.

Table 1: ***Descriptions of Some Common Human-Rater Errors and Biases***

Severity/ Leniency	Refers to a phenomenon in which raters make judgments on a common dimension, but some raters tend to consistently give high scores (leniency) while other raters tend to consistently give low scores (severity), thereby introducing systematic biases.
Scale Shrinkage	Occurs when human raters don't use the extreme categories on a scale.
Inconsistency	Occurs when raters are either judging erratically, or along different dimensions, because of their different understandings and interpretations of the rubric.
Halo Effect	Occurs when the rater's impression from one characteristic of an essay is generalized to the essay as a whole.
Stereotyping	Refers to the predetermined impression that human raters may have formed about a particular group that can influence their judgment of individuals in that group.
Perception Difference	Appears when immediately prior grading experiences influence a human rater's current grading judgments.
Rater Drift	Refers to the tendency for individual or groups of raters to apply inconsistent scoring criteria over time.

"Involving humans in grading such high volumes, especially in large-scale assessments like the GRE test, can be labor intensive, time consuming, and expensive."

It is also worth emphasizing that there has been relatively little published research on human-rater cognition (e.g., see Suto, Crisp, & Greatorex, 2008). Hence, what goes on in a rater's mind when passing judgment is not well known, particularly under operational scoring conditions. This lack of knowledge about the cognitive basis for human scoring could, in turn, undermine confidence in the validity of the automated scores produced by computers designed to emulate those human ratings. It is because of these known limitations of human scoring that consequential testing programs, such as those for admissions or licensure, typically use more than one human rater for each response, and adjudicate discrepancies between the raters if necessary.

Automated Scoring

Automated scoring has the potential to provide solutions to some of the obvious shortcomings in human essay scoring (e.g., rater drift). Today's state-of-the-art systems for computer-based scoring involve construct-relevant aggregation of quantifiable text features in order to evaluate the quality of an essay. These systems work exclusively with variables that can be extracted and combined mathematically. Humans, on the other hand, make holistic decisions under the influence of many interacting factors.

The primary strength of automated scoring compared to human scoring lies in its efficiency, absolute consistency in applying the same evaluation criteria across essay submissions and over time, as well as its ability to provide fine-grained, instantaneous feedback. Computers are neither influenced by external factors (e.g., deadlines) nor emotionally attached to an essay. Computers are not biased by their stereotypes or preconceptions of a group of examinees. Automated scoring can therefore achieve greater objectivity than human scoring (Williamson, Bejar, & Hone, 1999). Most automated scoring systems can generate nearly real-time performance feedback on various dimensions of writing. For example, ETS's *e-rater*® engine can provide feedback on grammar, word usage, mechanics, style, organization, and development of a written text (ETS, 2008). Similarly, Pearson's Intelligent Essay Assessor™ can provide feedback on six aspects of writing — ideas, organization, conventions, sentence fluency, word choice, and voice (Pearson Education, Inc., 2011). It would be quite difficult for human raters to offer such analytical feedback immediately for large numbers of essays.

Automated scoring systems are often able to evaluate essays across grade levels (e.g., the *e-rater* engine, Intelligent Essay Assessor, Vantage Learning's IntelliMetric®). Human graders, in contrast, are usually trained to focus on a certain grade range associated with a specific rubric and a set of tasks. Shifting a human rater to a new grade range may therefore require considerable retraining.

Further, some automated scoring systems are able to grade essays written in languages other than English (e.g., Intelligent Essay Assessor; Pearson Education, Inc., 2012). This capability could facilitate the scoring of tests that are translated into other languages for international administration, relieving the potential burden of recruiting and training a large pool of human graders for alternate language assessment. There

"Today's state-of-the-art systems for computer-based scoring involve construct-relevant aggregations of quantifiable text features in order to evaluate the quality of an essay. They work exclusively with variables that can be extracted and combined mathematically. Humans, on the other hand, make holistic decisions under the influence of many interacting factors."

are also automated scoring systems that are able to detect inauthentic authorship (e.g., IntelliMetric; Rudner, Garcia, & Welch, 2005), which human raters may not be able to do as readily as computers. It is worth noting that these alternate language and inauthentic authorship capabilities have not been widely researched. Still, these directions represent a potential path to improve upon human scoring.

Notwithstanding its strengths, it must be recognized that automated scoring systems generally evaluate relatively rudimentary text-production skills (Williamson et al., 2010), such as the use of subject-verb agreement evaluated by the grammar feature in the *e-rater* engine, and spelling and capitalization as evaluated by the mechanics feature. Current automated essay-scoring systems cannot directly assess some of the more cognitively demanding aspects of writing proficiency, such as audience awareness, argumentation, critical thinking, and creativity. The current systems are also not well positioned to evaluate the specific content of an essay, including the factual correctness of a claim. Moreover, these systems can only superficially evaluate the rhetorical writing style of a test taker, while trained human raters can appreciate and evaluate rhetorical style on a deeper level.

A related weakness of automated scoring is that these systems could potentially be manipulated by test takers seeking an unfair advantage. Examinees may, for example, use complicated words, use formulaic but logically incoherent language, or artificially increase the length of the essay to try and improve their scores. Powers, Burstein, Chodorow, Fowles, and Kukich (2001) conducted an experiment designed to "stump" an earlier version of the *e-rater* engine, and found that the engine was susceptible to such strategies. For example, it gave the highest score to a very long essay that contained 37 repetitions of several paragraphs while human raters gave the same text the lowest score possible. Although this particular issue had been resolved for the *e-rater* engine, as have some similar issues, unscrupulous examinees may still be able to improve their scores by using sophisticated words and writing extended, pre-memorized text. (Of course, even human raters may be influenced by some of these strategies.) As a consequence, it is important to be aware of the implications such manipulation can have for score validity and fairness.

A final weakness of automated scoring systems is that they are generally designed to "learn" the evaluation criteria by analyzing human-graded essays. This design implies that automated scoring systems could inherit not only positive qualities, but also any rating biases or other undesirable patterns of scoring present in the scores from human raters.

Table 2 summarizes the strengths and weaknesses of human and automated scoring as discussed above. The summary compares and contrasts the two methods from two aspects: measurement and logistical effort.

"Current automated essay-scoring systems cannot directly assess some of the more cognitively demanding aspects of writing proficiency such as audience awareness, argumentation, critical thinking, and creativity."

Table 2: ***A Summary of Strengths and Weaknesses in Human and Automated Scoring of Essays***

	Human Raters	Automated Systems
Potential Measurement Strengths	*Are able to:* • Comprehend the meaning of the text being graded • Make reasonable and logical judgments on the overall quality of the essay *Are able to incorporate as part of a holistic judgment:* • Artistic/ironic/rhetorical styles • Audience awareness • Content relevance (in depth) • Creativity • Critical thinking • Logic and argument quality • Factual correctness of content and claims	*Are able to assess:* • Surface-level content relevance • Development • Grammar • Mechanics • Organization • Plagiarism (some systems) • Limited aspects of style • Word usage *Are able to more efficiently (than humans) provide* (adapted from Williamson, et al., 1999): • Granularity (evaluate essays with detailed specifications with precision) • Objectivity (evaluate essays without being influenced by emotions and/or perceptions) • Consistency (apply exactly the same grading criteria to all submissions) • Reproducibility (an essay would receive exactly the same score over time and across occasions from automated scoring systems) • Tractability (the basis and reasoning of automated essay scores are explainable)
Potential Measurement Weaknesses	*Are subject to:* • Drift error • Halo effect • Inconsistency error • Subjectivity • Perception difference error • Severity error • Scale shrinkage error • Stereotyping error	*Are unlikely to:* • Have background knowledge • Assess creativity, logic, quality of ideas, unquantifiable features *And:* • Inherit biases/errors from human raters
Logistical Strengths		*Can allow:* • Quick re-scoring • Reduced cost (particularly in large-scale assessments) • Timely reporting including possibility of instantaneous feedback
Logistical Weaknesses	*Will require:* • Attention to basic human needs (e.g., housing, subsistence level) • Recruiting, training, calibration, and monitoring • Intensive direct labor and time	*Will require:* • Expensive system development • System maintenance and enhancement (indirect labor and time)

"When implementing a test, developers and program administrators can weigh the pros and cons and leverage the use of automated and human scoring methods according to their goals."

Practical Implications

The differences between human and automated scoring have important practical implications for how automated scoring can be implemented in a testing program, as well as for the validation of the automated scores.

Implications for Implementation

When implementing a test, developers and program administrators can weigh the pros and cons and leverage the use of automated and human scoring methods according to their goals. There are many ways to integrate an automated scoring system into a testing program. Figure 1 shows four common approaches, which vary from relying completely on automated scoring to relying less on automated scoring, including using the latter to verify human ratings.

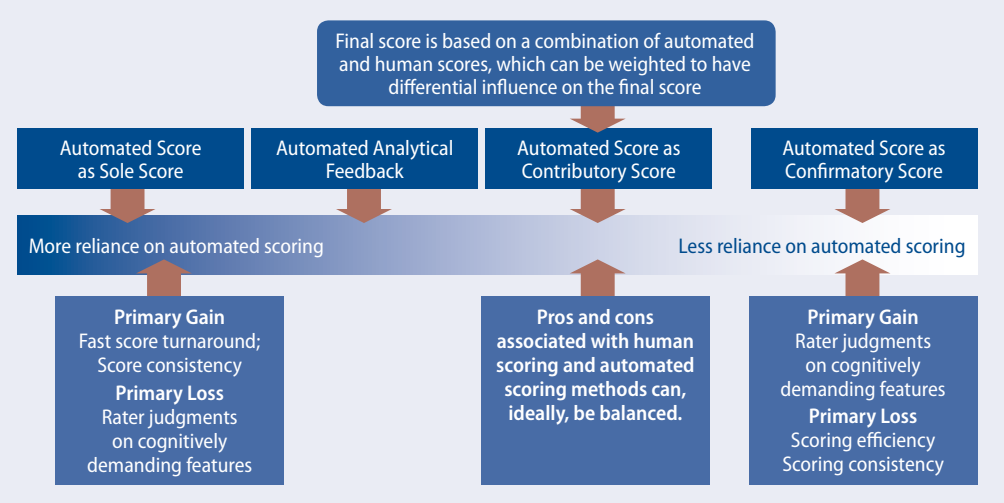
Use of automated scoring only

The approach that relies most on automated scoring capability reports the final test results from the automated scoring system alone, without any human scoring. This approach takes complete advantage of the benefits of automated scoring while sacrificing the strengths and capabilities associated with human grading.

This approach is used in large-scale, low-stakes tests such as ETS's *TOEFL*® Practice Online (TPO), College Board's ACCUPLACER®, and ACT's COMPASS®. In TPO, for example, the *e-rater* engine is the only essay grader; the examinees are primarily interested in practice opportunities with the format of the TOEFL iBT test and in getting feedback on their essay writing.

An approach that also relies primarily on automated scoring involves generating instant performance feedback. Current state-of-the-art technologies can already provide feedback on a number of dimensions (e.g., word sophistication, grammar). Fine-grained analytical comments such as those provided by MY Access!® (Vantage Learning), WriteToLearn® (Pearson Education, Inc.), and the *Criterion*® Online Writing Evaluation Service (ETS) have the potential to enhance students' daily educational experience.

Figure 1: ***Common Implementation Approaches to Automated Scoring in a Testing Program***



"Only when the system's procedures for generating scores are transparent can score users and the scientific community associated with those users fully evaluate the system's usefulness and potential liabilities."

While automated scoring can defensibly be used as the sole evaluation mechanism in low-stakes settings, the current state of the art is not sufficiently mature to endorse automated essay scoring as the sole score in assessments linked to high-stakes decisions. The following three conditions need to be met before automated scoring is applied in high-stakes environments as the only scoring method for essays. ETS and other institutions are doing substantive research and development toward meeting these goals.

- The internal mechanism for an automated scoring system is reasonably transparent, especially in cases where the automated scores are used to contribute to high-stakes decisions. Only when the system's procedures for generating scores are transparent can score users and the scientific community associated with those users fully evaluate the system's usefulness and potential liabilities.
- A sufficiently broad base of validity evidence is collected at the early stage of the implementation so that the deployment of the automated scoring system can be justified.
- A quality-control system is in place so that aberrant performance on the part of the automated scoring system can be detected in time to prevent reporting incorrect results.

Use of automated scoring in conjunction with human rating

There are at least two alternatives that rely less on automated scoring compared to the approach that uses automated scoring exclusively. Both apply automated scoring in tandem with human scoring. In the first alternative, the final essay grade is based on a combination of human ratings and automated scores (additional human raters are brought in if the difference between the human and automated rating is larger than a preset threshold). One advantage of this approach is that developers and users, such as testing program administrators and educators, can balance the influence of the two methods on the final score by managing the weights. For example, automated scores could be given equal weight to human ratings on the final scores (e.g., 1/2 human + 1/2 machine), or they could be assigned a higher or lower weight than human ratings (e.g., 1/3 human + 2/3 machine or 2/3 human + 1/3 machine). Ideally, the various weighting strategies should be flexible enough to allow the developers and users to achieve a balance that matches their comfort level with each scoring method (which, of course, should be supported by validity evidence). Applying automated scoring in this way could reduce the degree of human-generated inconsistency in the final scores.

In the second alternative, automated scoring is used as a quality-control mechanism for human scoring, with additional graders called in to adjudicate when human ratings differ from automated scores by more than a predetermined threshold. However, only the human scores are used for score reporting; the automated scores are used solely to determine whether a single human score is sufficient. This approach ensures that human raters' judgments are used for evaluating cognitively demanding writing skills, while sacrificing few of the benefits of automated scoring.

"Human raters typically are used not only as a development target, but also as an evaluation criterion. Agreement in itself, however, is not enough to support the use of automated scores, particularly if evidence supporting the validity of human ratings is insufficient."

There are successful examples of using the *e-rater* engine in ETS testing products representing each of these alternatives. The *e-rater* engine contributes to the final TOEFL iBT essay score in conjunction with human ratings, and it is used to confirm human ratings in the GRE Analytical Writing assessment. For both programs, decisions on implementation were made in collaboration with program administrators and based on substantive research evidence addressing both human and automated scoring methods.

Implications for Validation

Human ratings are often used as a development target, as well as evaluation criteria, for automated scoring.

Human rating as a development target

The scores generated by automated systems are usually modeled to predict operational human ratings, but there is an underlying assumption that those ratings are a valid development target. If not, errors and biases associated with human ratings could propagate to the automated scoring system.[2] Aside from gathering data to support the validity of operational human ratings, approaches to deal with this issue might include modeling the ratings gathered under conditions where the raters are under less pressure, and constructing scoring models using experts' judgments of the construct relevance and importance of different text features.

Human rating as an evaluation criterion

Human raters are typically used not only as a development target, but also as an evaluation criterion. Agreement in itself, however, is not enough to support the use of automated scores, particularly if evidence supporting the validity of human ratings is insufficient (e.g., in terms of their scoring process, resistance to manipulation). Even if human ratings were an "ultimate criterion," automated scores rarely, if ever, agree with them perfectly, leaving considerable room for the two methods to measure different attributes. Therefore, validation should be based on a broad collection of evidence, including an investigation of construct relevance and coverage for both automated and human scores (Bennett, 2011; Williamson, Xi, & Breyer, 2012). Automated essay scores should, for example, correlate highly with external measures of the same construct (i.e., writing competency) and more weakly with measures of different constructs (e.g., proficiency in math or chemistry). When automated scoring is used on heterogeneous populations, evidence should be collected to ensure that both the automated and human scores carry the same meaning from one population group to another. One way to do so is to examine the invariance of the human-machine agreement across population groups. A lack of invariance would raise questions as to the validity and fairness of both methods. Whether human and automated scores have the same correlational pattern with other measures can also be investigated. In

[2] This concept was referred to as a "first-order validity argument" in Bejar (2012). More specifically, appraisal of the human ratings is a prerequisite for all subsequent assumptions and inferences, and the omission of such an appraisal would lead one to question the subsequent claims and conclusions.

any event, developers and users of testing programs need to be aware of the potential consequences of any differences in the score meaning of human or automated systems for population groups.

Looking Into the Future

In 1999, the *e-rater* engine became the first automated scoring engine used operationally in a large-scale testing program associated with high-stakes decisions (i.e., the GMAT®). Since that time, much research has been conducted by ETS and other organizations, continuously advancing the state of the art in order to strengthen the validity of automated scoring, and to find better ways of blending the two scoring methods. One of the goals of this research is to gain a deeper understanding of the human raters' cognition and behavior. It is clear that our lack of knowledge about human cognition in essay scoring limits our ability to directly compare human and automated scoring. This gap in the research literature is one factor behind an ETS study in which scientists use eye-tracking technology to investigate what human raters attend to or ignore when grading essays. This research should increase our understanding of why human raters agree (or disagree) with one another, as well as why humans and machines agree (or disagree).

Closing Statement

Advances in artificial intelligence technologies have made machine scoring of essays a realistic option. Research and practical experience suggest that the technique is promising for various testing purposes, and that it will be used more widely in educational assessments in the near future. However, it is important to understand the fundamental differences between automated and human scoring and be aware of the consequences of choices in scoring methods and their implementation. Knowing their strengths and weaknesses allows testing program directors and policymakers to make strategic decisions about the right balance of methods for particular testing purposes. Automated scoring can — when carefully deployed — contribute to the efficient generation and delivery of high-quality essay scores, and it has the potential to improve writing education in K–12, as well as in higher education, as the capability becomes more mature.

Acknowledgments

The author would like to thank David Williamson, Randy Bennett, Michael Heilman, Keelan Evanini, Isaac Bejar, and the editor of *R&D Connections* for their helpful comments on earlier versions of this essay. The author is solely responsible for any errors that may remain. Any opinions expressed in this essay are those of the author, and not necessarily those of Educational Testing Service.

References

Baldwin, D., Fowles, M., & Livingston, S. (2005). Guidelines for constructed-responses and other performance assessments. Retrieved from **http://www.ets.org/Media/About_ETS/pdf/8561_ConstructedResponse_guidelines.pdf**

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

Bejar, I. I. (2012). Rater cognition: implications for validity. *Educational Measurement: Issues and Practice*, *31*(3), 2–9. **http://onlinelibrary.wiley.com/doi/10.1111/j.1745-3992.2012.00238.x/abstract**

Bennett, R. E. (2011). *Automated scoring of constructed response literacy and mathematics items*. Retrieved from **http://www.ets.org/s/k12/pdf/k12_commonassess_automated_scoring_math.pdf**

Educational Testing Service. (2008). Criterion℠ *online writing evaluation service*. Retrieved from **http://www.ets.org/s/criterion/pdf/9286_CriterionBrochure.pdf**

Educational Testing Service. (2013). GRE® *program sees volume surge in 2012 peak testing period*. Retrieved from **http://www.ets.org/newsroom/news_releases/gre_volume_surge**

Pearson Education, Inc. (2011). *Demonstrating reading & writing performance gains*. Retrieved from **http://www.writetolearn.net/downloads/WTL_EfficacyReport.pdf**

Pearson Education, Inc. (2012). *Intelligent Essay Assessor™ (IEA) fact sheet*. Retrieved from **http://kt.pearsonassessments.com/download/IEA-FactSheet-20100401.pdf**

Powers, D. E., Burstein, J. C., Chodorow, M., Fowles, M. E., & Kukich, K. (2001). *Stumping* e-rater*: challenging the validity of automated essay scoring*. GRE Board Professional Report No. 98-08bP, ETS Research Report 01-03. **http://www.ets.org/Media/Research/pdf/RR-01-03-Powers.pdf**

Rudner, L. M., Garcia, V., & Welch, C. (2005). *An evaluation of IntelliMetric™ essay scoring system using responses to GMAT® AWA prompts*. Retrieved from **http://www.gmac.com/~/media/Files/gmac/Research/research-report-series/RR0508_IntelliMetricAWA**

Suto, I., Crisp, V., & Greatorex, J. (2008). Investigating the judgmental marking process: an overview of our recent research. *Research Matters*, *5*, 6–9.

Williamson, D. M., Bejar, I. I., & Hone, A. S. (1999). Mental model comparison of automated and human scoring. *Journal of Educational Measurement*, *36*, 158–184.

Williamson, D. M., Bennett, R. E., Lazer, S., Bernstein, J., Foltz, P. W., Landauer, T. K., & Sweeney, K. (2010). *Automated scoring for the assessment of common core standards*. Retrieved from **http://research.collegeboard.org/sites/default/files/publications/2012/8/ccss-2010-5-automated-scoring-assessment-common-core-standards.pdf**

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice*, *31*, 2–13. **http://onlinelibrary.wiley.com/doi/10.1111/j.1745-3992.2011.00223.x/abstract**

R&D Connections is published by

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

Editor: Hans Sandberg
Copy Editor: Eileen Kerrigan
Layout Design: Sally Acquaviva

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

Follow ETS Research on Twitter® **(@ETSresearch)**