ACT Research & Policy

Issue Brief

2015

Reviewing Your Options: The Case for Using Multiple-Choice Test Items

Michelle Croft, Gretchen Guffy, and Dan Vitale

As the emphasis on college and career readiness standards has given greater urgency to debates over the value of high-stakes testing, multiple-choice test questions have come in for a new round of criticism. Critics contend, for example, that multiple-choice questions cannot measure higher-order thinking skills or reflect real-world problem solving.[1]

However, when constructed well, multiple-choice (sometimes called selected-response) questions can and do efficiently assess students' higher-order thinking skills and reflect their real-world problem solving skills, and are an important part of an assessment system that includes a variety of question formats and types of assessments.[2] This issue brief identifies the many benefits of multiple-choice questions and offers policy recommendations to support the appropriate use of multiple-choice questions.

Why Use Multiple-Choice Test Questions?

Multiple-choice questions have many desirable features, which is why ACT invests time and resources in their development. We detail a number of these features in the following sections.

Content Coverage and Testing Time

One of the chief advantages of multiple-choice questions is their efficiency. Multiple-choice questions can cover a broader range of content than either constructed-response tasks (tasks that require a longer response, such as an essay) or performance-based tasks (tasks that require the test taker to complete a task, such as a science experiment) in less testing time.[3] This is particularly important when assessments must measure a wide range of content standards without taking up too much classroom time.[4]

Multipurpose

Some people believe, incorrectly, that multiple-choice questions only measure recall, or the ability of a test taker to recognize or retrieve a fact or some other bit of discrete information.[5] While multiple-choice questions may be well suited to testing this type of retrieval, they are by no means limited to it.[6]

Multiple-choice questions can also assess higher-order thinking skills.[7] For instance, posing a scenario followed by several questions that require test takers to apply what they have learned in the scenario is one way of measuring higher-order thinking skills.[8] Likewise, asking test

Michelle Croft, PhD, JD, is a principal research associate in the Office of Policy, Advocacy, and Government Relations.

Gretchen Guffy, MPP, is director of policy in the Office of Policy, Advocacy, and Government Relations.

Dan Vitale is senior associate in the Office of Policy Advocacy, and Government Relations.

Email research.policy@act.org for more information.

takers to recognize a pattern and use the pattern to solve the problem, as in the sample problem in figure 1, goes beyond measuring recall and requires students to demonstrate sophisticated problem-solving skills.

Reliability and Validity

Scores from well-constructed multiple-choice questions are generally highly reliable. Reliability is the consistency of test scores either internally (that is, the questions are generally measuring the same construct) or from one test administration to the next.[9] The more questions that are included on an assessment, the higher its reliability usually is.[10] Since there are typically more questions on a multiple-choice assessment than on a constructed-response or performance-based assessment, multiple-choice assessments generally have higher reliability coefficients.[11] Further, constructed-response and performance-based assessments are more susceptible to increased measurement error due to "person by task interaction," where the person has the knowledge to correctly respond to the task but "may react to the specific context or other extraneous characteristics of the task."[12] In practical terms, this means that test takers are more likely to receive similar scores if they retake a multiple-choice test than if they retake a constructed-response or performance-based test.

Another advantage of multiple-choice tests is validity. Validity is the degree to which evidence and theory support the interpretations represented by the test scores, as well as the degree to which those interpretations are dependent on the proposed uses of the scores.[13] Because multiple-choice tests can cover a broad range of content in a relatively short amount of time, potentially covering more standards, they typically have more validity evidence; we should therefore have more confidence in the score interpretations.[14]

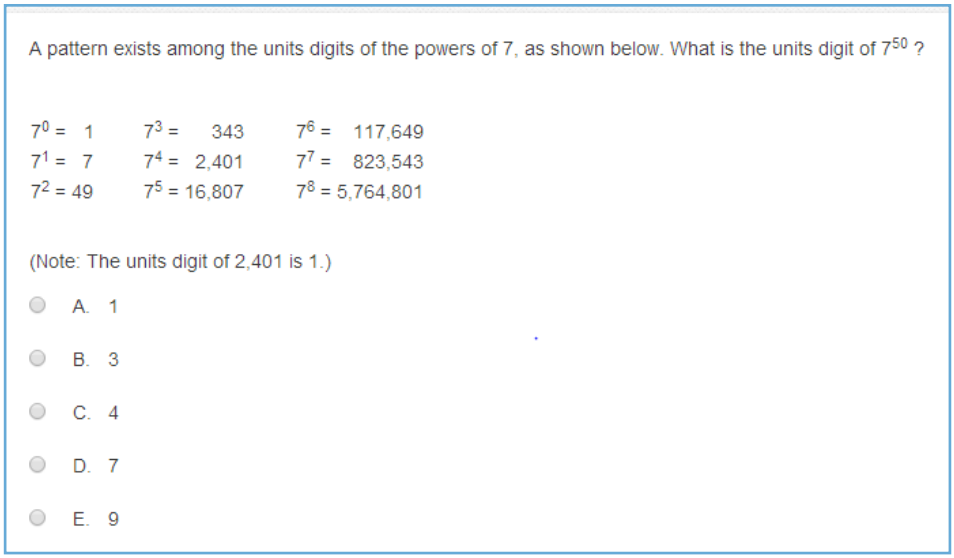

Figure 1. Sample ACT Aspire® Mathematics test question for grades 5 to early high school (grades 9–10). Answer option E (9) is the correct answer.

Despite broad content coverage, critics contend that multiple-choice tests cannot represent what is necessary to demonstrate college and career readiness.[15] Advocates of constructed-response and performance-based assessments emphasize these assessments' potential to "enhance" the validity of scores by requiring test takers to provide rather than select a response.[16] Providing a response increases what measurement experts refer to as "face validity," or what the test "appears superficially to measure."[17] For example, an essay test requires students to write an essay, and a test that requires a student to perform an experiment insists on the student's performing the experiment in order to be scored on the task. However, writing an essay or performing an experiment does not necessarily mean a student's problem-solving and higher-order thinking skills are being tested; knowing that requires further validity research.[18] "It is all too easy to think of higher-order skills as involving only difficult subject matter as, for example, learning calculus," observes one report on *The Nation's Report Card*. "Yet one can memorize the formulas for derivatives just as easily as those for computing areas of various geometric shapes, while remaining equally confused about the overall goals of both activities."[19] Gathering validity evidence is important regardless of assessment format.[20]

Student Engagement

Students indicate that they like the multiple-choice format.[21] One reason is that they perceive the item type as easier,[22] likely due to familiarity, but another is that the presence of answer choices provides students with feedback, helping them know if they are interpreting the question correctly.[23] For instance, in interviews with students, researchers found that some students had determined the correct answer to constructed-response tasks but had written nothing because they were unsure whether they had correctly understood the question.[24]

This is an issue because studies have found higher omission rates—that is, the rates at which students omit or "skip over" a test

item—for constructed-response tasks than for multiple-choice questions.[25] In a study of the 1992 National Assessment of Educational Progress (NAEP) reading assessment and the 1996 NAEP math assessment, researchers found that constructed-response tasks were omitted by about 8% of students.[26] For reading, the maximum omission rate was 18% for constructed response but only 4% for multiple choice.[27] For math, the difference was even greater, with maximum omission rates of 25% for constructed response and 5% for multiple choice.[28] A 2007 study in Ohio found similar omission rates. In their third-grade sample, omissions ranged from a low of 1.3% in mathematics to a high of 32% in reading for constructed-response tasks, whereas none of the multiple-choice questions had an omission rate higher than 1%.[29]

The omission rate issue is also potentially problematic with respect to the race/ethnicity of test takers: because Hispanic and African American students are more likely to omit responses than White students, more Hispanic and African American test takers may be disadvantaged when taking assessments, such as constructed-response tests, that generate higher omission rates.[30]

Guessing

One reason there are fewer omissions on multiple-choice questions than on constructed-response or performance-based tasks is that multiple-choice questions present response options to the test taker. A test taker who does not know the answer is still able to guess.

There are ways to reduce guessing. One way is to have a common stem and response options with multiple parts, as in the eighth-grade ACT Aspire Mathematics test question in figure 2. Students are given a prompt ("Click on all of the irrational numbers below") for eight different square roots. They must respond "yes" or "no" by either clicking or not clicking on each of the square roots. Guessing is reduced for this question because students are required to identify all four irrational numbers ($\sqrt{2}$, $\sqrt{8}$, $\sqrt{32}$, and $\sqrt{128}$) to receive credit.

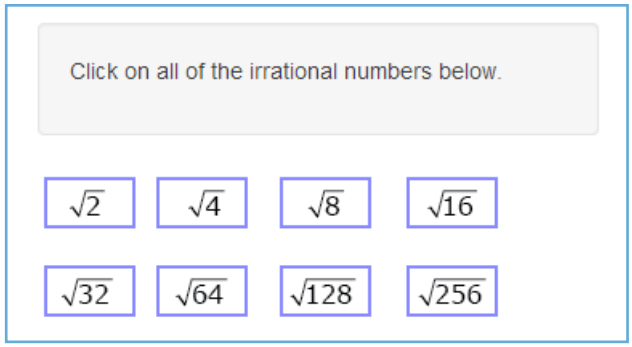

Figure 2. Sample ACT Aspire Mathematics test question for eighth grade

Cost

Multiple-choice questions have an advantage over other item types in that they can be scored automatically. Automatic scoring is much more economical than other kinds of scoring. After the passage of the No Child Left Behind Act of 2001, the General Accounting Office (GAO) estimated tests, the development, administration, and scoring of different types of assessments. For solely multiple-choice assessments, the cost was $1.90 billion. For a mixture of multiple-choice and open-ended items, the cost rose to $5.31 billion.[31] Scoring accounted for the bulk of the cost difference. Indeed, the GAO estimated that the percentage of the total cost of testing related to administration, scoring, and reporting was 65% for multiple-choice tests and 86% for tests with both multiple-choice and open-ended items. Similarly, the costs of performance-based assessments are even higher. In the early 1990s, the GAO found that a solely performance-based assessment would cost $18 more per student than a multiple-choice test.[32]

As innovations in artificial intelligence scoring help improve our ability to use computers to score open-ended tasks, this cost discrepancy may diminish. However, early evidence suggests that computer-based scoring should only be used as a "second reader," necessitating the continued costs of having at least one human scorer as well as the costs of rater training to ensure a high level of consistency in scoring.[33]

Besides the financial cost of scoring open-ended tasks, there is a cost of time. Consider the Maryland School Performance Assessment Program, a highly regarded performance-based assessment in the 1990s. Its language arts assessment occurred in five sessions over five consecutive school days, with sixty or ninety minutes of testing per session,[34] while its mathematics assessment was administered in three one-hour sessions on three consecutive school days either preceding or following the five days of the language arts assessment.[35] All in all, Maryland students were tested for at least one hour per day over eight consecutive school days, for a minimum of eight hours total. Compare this to the ACT Aspire reading, mathematics, and science tests within ACT Aspire, which take approximately an hour each, and the English and writing tests, which take approximately thirty minutes each, for a total of approximately four hours.[36]

Policy Recommendations

The use of multiple-choice questions in assessment systems efficiently and cost-effectively returns valuable information about what test takers know and can do. We ask policymakers and educators to consider the following policy recommendations when selecting item formats and assessments:

1. **Unless the content standards explicitly require the student to "perform" a task (e.g., conduct an experiment), consider using a selected-response or multiple-choice question.** Assessment results provide a valuable resource to schools, but assessments do take instructional time. Selection of an assessment and the item types used should be purposeful to determine what types of skills need to be measured and the most effective ways to measure them. Multiple-choice questions are an efficient and cost-effective way to measure what a student knows and can do, including higher-order thinking skills.
2. **Regardless of item type, anyone developing a test should carefully evaluate validity evidence to determine whether items are measuring higher-order thinking skills.** Item format alone does not dictate whether an item is measuring higher-order thinking. Instead, there must be validity evidence to support labeling an item as measuring higher-order thinking skills.
3. **When adopting state or district assessment systems, ensure that all test items—multiple choice, constructed response, and performance based—have been rigorously tested and measure a broad range of content.** States and districts have many options when adopting assessment systems. They should ensure that the assessment system they choose is aligned with content standards and supported by adequate reliability and validity evidence. ■

Appendix: Multiple-Choice Test Item Development for the ACT

There are five main steps in the multiple-choice item development process for the ACT. Each step encompasses numerous processes and reviews: for an item to make it to an operational test form (a form in which a student's performance on the item contributes to the student's test score), it must pass as many as sixteen stages of review.

Step 1: Item Writing

Item writers (who are content specialists; many are active teachers) are given a guide specific to the content area for which they have been engaged to develop items. The guide provides the test specifications, content and style requirements (i.e., criteria for accuracy, word count, item classification, format, and language), and examples of acceptable items. ACT staff edit the items to ensure that each item meets these requirements.

Step 2: External Reviews

External content and fairness experts review items for accuracy, grade-level appropriateness, educational importance, and fairness to all test takers. ACT staff then further edit each item as required for it to conform to the expert feedback.

Step 3: Tryout and Statistical Analysis

Items are piloted in unscored sections of the test for which they were developed.[37] ACT staff then conduct statistical analyses on each of the piloted items to determine whether the items contain statistical irregularities. The analyses help to identify items that are too easy, too difficult, or that do not differentiate between high- and low-performing test takers.[38] ACT staff also review all items flagged for statistical irregularities to determine whether an item can be revised for a subsequent tryout or must be discarded as unusable.

Step 4: Item Pool

If an item successfully passes tryout, it can be placed in an item pool for use on an operational test form. Items for new forms are selected from a pool based on content criteria and statistical properties. The statistical goal is to create a form that is similar to prior forms with respect to its average difficulty[39] and its ability to effectively differentiate among students at different performance levels.[40] Items are also selected to ensure that students will have sufficient time to complete the whole test.[41]

Step 5: Additional External Reviews

ACT staff review the test form as a whole for content accuracy and style. New sets of external content and fairness experts then review the form. Based on feedback from the experts, necessary changes are made to the test forms before they can be administered operationally.

Notes

1 For example, in a recent white paper Parsi and Darling-Hammond state that multiple-choice tests do not offer any information about students' thinking or reasoning, their misconceptions, or ability to express their ideas. Ace Parsi and Linda Darling-Hammond, *Performance Assessments: How State Policy Can Advance Assessments for 21st Century Learning* (National Association of State Boards of Education and Stanford Center for Opportunity Policy in Education, 2015), http://www.nasbe.org/wp-content/uploads/Parsi-LDH-Performance-Assessment_Jan2015.pdf.

2 It is important to note that "assessment formats should vary according to the type of standards that need to be measured," as "multiple measures can be used to offer more comprehensive evaluations of student achievement, from multiple-choice and constructed-response assessments to performance tasks and project-based learning." ACT, *Policy Platform K–12* (Iowa City, IA: ACT, December 2014), http://www.act.org/policyplatforms/pdf/k-12-online.pdf. See the appendix for a discussion of multiple-choice item test development for the ACT® test.

3 Thomas M. Haladyna, *Developing and Validating Multiple-Choice Items* (Mahwah, NJ: Lawrence Erlbaum Associates, 1999); Ruth A. Childs and Andrew P. Jaciw, "Matrix Sampling of Test Items," *ERIC Digest* (ERIC Clearinghouse on Assessment and Evaluation, 2003), http://www.ericdigests.org/2005-1/matrix.htm. Also, both teachers and students have commented on the length of extended response tasks in interviews about SBAC. One student memorably stated, "It was soooooo long." Nancy Doorey, *Smarter Balanced "Tests of the Test" Successful: Field Test Provides Clear Path Forward* (Smarter Balanced Assessment Consortium, 2014), http://www.smarterbalanced.org/wordpress/wp-content/uploads/2014/10/SmarterBalanced_FieldTest_Report.pdf.

4 US Department of Education, *NCLB Standards and Assessments Non-Regulatory Guidance* (Washington, DC: US Department of Education, 2013), http://www2.ed.gov/policy/elsec/guid/saaguidance03.doc. US Department of Education, "Overview Information; Race to the Top Fund Assessment Program; Notice Inviting Applications for New Awards for Fiscal Year (FY) 2010," *Federal Register* 75 (April 9, 2010), 18171, http://www.gpo.gov/fdsys/pkg/FR-2010-04-09/pdf/2010-8176.pdf.

5 See "Bloom's Taxonomy," Vanderbilt University Center for Teaching, http://cft.vanderbilt.edu/guides-sub-pages/blooms-taxonomy; and "Depth of Knowledge," New York City Department of Education, http://schools.nyc.gov/Academics/CommonCoreLibrary/ProfessionalLearning/DOK/default.htm. Norman L. Webb, "Depth-of-Knowledge Levels for Four Content Areas," last modified March 28, 2002, http://facstaff.wcer.wisc.edu/normw/All%20content%20areas%20%20DOK%20levels%2032802.doc.

6 By the same token, it is not necessarily true that item types other than multiple choice always assess higher-order thinking skills. For example, constructed-response tasks require test takers to provide an answer, but writing a response does not necessarily mean that higher-order thinking skills are involved. Tasks could be written to only require factual recall. For example, asking students to provide, rather than select, the answer to the factual question "What is 7 minus 3?" is still considered a constructed-response task.

7 Richard P. Phelps, *Standardized Testing* (New York, NY: Peter Lang Publishing, 2007), 92–93.

8 Karyn Woodford and Peter Bancroft, "Using Multiple Choice Questions Effectively in Information Technology Education," in R. Atkinson, C. McBeath, D. Jonas-Dwyer, and R. Phillips (Eds.), *Beyond the Comfort Zone: Proceedings of the 21st ASCILITE Conference* (Perth, Australia: 2004), 948–955, http://www.ascilite.org.au/conferences/perth04/procs/pdf/woodford.pdf. Also see Teaching Effectiveness TEP Program, "Techniques for Writing Multiple-Choice That Demand Critical Thinking," http://tep.uoregon.edu/resources/assessment/multiplechoicequestions/sometechniques.html.

9 American Educational Research Association, American Psychological Association, and the National Council on Measurement in Education, *Standards for Educational and Psychological Testing* (Washington, DC: American Educational Research Association, 2014).

10 Ibid., 38.

11 See a discussion of "task sampling variability" in Brian Stecher, *Performance Assessment in an Era of Standards-Based Educational Accountability* (Stanford, CA: Stanford University, Stanford Center for Opportunity Policy in Education, 2010), https://scale.stanford.edu/system/files/performance-assessment-era-standards-based-educational-accountability.pdf.

12 Committee for the Workshop on Alternatives in Assessing Adult Education and Literacy Programs, National Research Council, Division of Behavioral and Social Sciences, and National Research Council, *Performance Assessments for Adult Education* (Washington, DC: National Academies Press, 2002), 45, citing R. L. Brennan & E. G. Johnson, "Generalizability of Performance Assessments," *Educational Measurement Issues and Practice* 14, no. 4 (1995): 9–12. Charlene G. Tucker, *Psychometric Considerations for Performance Assessment with Implications for Policy and Practice* (Princeton, NJ: K-12 Center at ETS, 2015).

13 Phelps, *Standardized Testing*, 87.

14 Paul M. La Marca, "Alignment of Standards and Assessments as an Accountability Criterion," *Practical Assessment, Research, & Evaluation* 7, no. 21 (2001), http://pareonline.net/getvn.asp?v=7&n=21.

15 See Linda Darling-Hammond, "Beyond the Bubble Test: Why We Need Performance Assessments," *Education Week*, July 9, 2014, http://blogs.edweek.org/edweek/education_futures/2014/07/beyond_the_bubble_test_why_we_need_performance_assessments.html.

16 Robert I. Linn, Eva L. Baker, and Stephen B. Dunbar, "Complex, Performance-Based Assessment: Expectations and Validation Criteria," CSE Technical Report 331 (Boulder, CO: University of Colorado, 1991), http://www.cse.ucla.edu/products/Reports/TECH331.pdf.

17 William A. Mehrens, "Using Performance Assessment for Accountability Purposes," *Educational Measurement: Issues and Practice* 11, no. 1 (1992): 3–9, citing A. Anastasi. *Psychological Testing*, 6th ed. (New York, NY: Macmillan, 1998).

18 Linn, Baker, and Dunbar, "Complex, Performance-Based Assessment."

19 National Academy of Education, *Commentary by the National Academy of Education on the Nation's Report Card* (Cambridge, MA: National Academy of Education, 1987), 54.

20 A psychometric review of the Maryland School Performance Assessment Program commented that there was no direct evidence the MSPA was measuring higher-order thinking skills and that it was confounding writing with higher-order thinking skills. Ronald K. Hambleton, James Impara, William Mehrens, Barbara S. Plake, Mary J. Pitoniak, and April L. Zenisky, *Psychometric Review of the Maryland School Performance Assessment Program* (2000), http://marces.org/mdarch/pdf/msde000003.pdf.

21 Doorey, *Smarter Balanced "Tests of the Test" Successful*.

22 Ibid.

23 US Department of Education, *NAEP Validity Studies*.

24 Ibid.

25 Koretz, Lewis, Skewes-Cox, and Burstein, examining the 1990 NAEP mathematics assessment blocks, found that almost all of the items that had omit rates above .10 were constructed-response tasks. However, the omit rates also appeared to be due to the difficulty of the item, where those that were omitted were "atypically difficult." Daniel Koretz, Elizabeth Lewis, Tom Skewes-Cox, and Leigh Burstein, "Omitted and Non-Reached Items in Mathematics in the 1990 National Assessment of Educational Progress," CSES Technical Report 357 (Los Angeles, CA: Center for Research on Evaluation, Standards, and Student Testing, 1993), http://www.cse.ucla.edu/products/Reports/TECH357.pdf.

26 US Department of Education, National Center for Education Statistics, *NAEP Validity Studies: An Investigation of Why Students Do Not Respond to Questions*, NCES 2003-12 (Washington, DC: US Department of Education, 2003), http://nces.ed.gov/pubs2003/200312.pdf.

27 US Department of Education, *NAEP Validity Studies.*

28 Ibid.

29 Liz Hollingworth, Jonathan J. Beard, and Thomas P. Proctor, "An Investigation of Item Type in a Standards-Based Assessment," *Practical Assessment Research & Evaluation* 12, no. 18 (2007), http://pareonline.net/getvn.asp?v=12&n=18. The authors also noted problems with the constructed response where student handwriting was illegible or answers were not written in English. For both of those cases, items were not able to be scored.

30 US Department of Education, *NAEP Validity Studies,* citing S. Swinton, *Differential Response Rates to Open-Ended and Multiple-Choice NAEP Items by Ethnic Groups* (Princeton, NJ: Educational Testing Service, 1991) and Daming Zhu and Tony D. Thompson, *Gender and Ethnic Differences in Tendencies to Omit Responses on Multiple-Choice Tests Using Number-Right Scoring*, paper presented at the annual meeting of the American Educational Research Association, April 18–22, 1995, ERIC Document Reproduction Service No. ED 382 689. Studies examining differential item functioning, item type, and race/ethnicity have found inconsistent results. One study found that the use of constructed-response itmes provides an advantage for white test takers compared to African American test takers. Robert W. Lissitz, Xiaodong Hou, and Sharon Cadman Slater, "The Contribution of Constructed Response Items to Large Scale Assessment: Measuring and Understanding their Impact," *Journal of Applied Testing Technology* 13, no. 3 (2012). Another found that constructed-response items favor African American or Hispanic students compared to White students. Catherine S. Taylor and Yoonsun Lee, "Ethnic DIF in Reading Tests with Mixed Item Formats," *Educational Assessment* 16, no. 1 (2011): 35–68.

31 United States General Accounting Office, *Characteristics of Tests Will Influence Expenses; Information Sharing May Help States Realize Efficiencies* (Washington, DC: GAO, 2003), http://www.gao.gov/new.items/d03389.pdf.

32 For a summary of research related to performance-assessment development and scoring costs, see also Lawrence O. Picus, Frank Adamson, William Montague, and Margaret Owens, *A New Conceptual Framework for Analyzing the Costs of Performance Assessment* (Stanford Center for Opportunity in Education, 2010), https://scale.stanford.edu/system/files/new-conceptual-framework-analyzing-costs-performance-assessment.pdf.

33 Caralee J. Adams, "Essay-Grading Software Seen as Time-Saving Too," *Education Week*, March 10, 2014, http://www.edweek.org/ew/articles/2014/03/13/25essay-grader.h33.html.

34 Wendy M. Yen and Steven Ferrara, "The Maryland School Performance Assessment Program: Performance Assessment with Psychometric Quality Suitable for High Stakes Usage," *Educational and Psychological Measurement* 57 (1997).

35 Ibid.

36 ACT, *Summative Assessment Technical Bulletin #1* (Iowa City, IA: ACT, 2014), http://www.discoveractaspire.org/pdf/2014_ACT-AspireTechnicalBulletin1.pdf.

37 To achieve a representative sample, the piloted items are placed in nationally administered test forms.

38 We separate students into low-, medium-, and high-performing groups and then calculate biserial and point-biserial correlation coefficients between each item score (correct/incorrect) and the total score on the corresponding test of the regular test form.

39 For example, the target mean item difficulty is about .58 for the ACT, with a range of difficulties from approximately .20 to .89.

40 The statistic used to measure how well an item differentiates among test takers is called the discrimination index. The ACT uses the biserial correlation as a measure of discrimination and targets items with a correlation of .20 or higher.

41 The item completion rate looks at the average proportion of students who completed each of the last five items.