A Guide to Computer Adaptive Testing Systems

Written by Tim Davey, Educational Testing Service
for
Technical Issues in Large-Scale Assessment (TILSA)
State Collaborative on Assessment and Student Standards (SCASS)
November 2011

THE COUNCIL OF CHIEF STATE SCHOOL OFFICERS

The Council of Chief State School Officers (CCSSO) is a nonpartisan, nationwide, nonprofit organization of public officials who head departments of elementary and secondary education in the states, the District of Columbia, the Department of Defense Education Activity, and five U.S. extra-state jurisdictions. CCSSO provides leadership, advocacy, and technical assistance on major educational issues. The Council seeks member consensus on major educational issues and expresses their views to civic and professional organizations, federal agencies, Congress, and the public.

A Guide to Computer Adaptive Testing Systems

COUNCIL OF CHIEF STATE SCHOOL OFFICERS

Christopher Koch (Illinois), President

Gene Wilhoit, Executive Director

Written by:
Tim Davey, Educational Testing Service
for
Technical Issues in Large-Scale Assessment (TILSA)
State Collaborative on Assessment and Student Standards (SCASS)

Council of Chief State School Officers

One Massachusetts Avenue, NW, Suite 700

Washington, DC 20001-1431

Phone (202) 336-7000

Fax (202) 408-8072

www.ccsso.org

OVERVIEW

Some brand names are used generically to describe an entire class of products that perform the same function. *Kleenex, Xerox, Thermos,* and *Band-Aid* are good examples. The term "*computerized adaptive testing*" (CAT) is similar in that it is often applied uniformly across a diverse family of testing methods. Although the various members of this family share similar goals and employ similar means, they differ in important ways that can significantly impact how tests perform when delivered. This guide is intended to provide those responsible for procuring CAT systems with questions they can use to tease out differences across competing CAT delivery systems, and so make better-informed decisions.

The scope of this guide is largely limited to differences across systems that can affect the quality, comparability, and usefulness of the test scores that a system produces. The substance of what is measured (the format and nature of the test questions) and test delivery issues (testing sites and test presentation software and hardware) will be considered only to the extent that they directly impact score properties.

This guide is organized around the elements common to nearly all varieties of adaptive testing. To orient the reader to this framework, the guide begins with a brief introduction to adaptive testing methods and the sometimes arcane vocabulary that has grown up around those methods.

ADAPTIVE TESTING

The basic principle behind adaptive testing is simple: avoid asking questions that are much too difficult or much too easy for the student being tested. Because we are fairly sure (but not certain!) that able students will answer easy items correctly and that struggling students will stumble on hard questions, relatively little is learned from such responses. Much more is learned by administering questions that challenge, but don't overwhelm, the student. Properly identifying and then presenting these questions is the goal of every adaptive test.

Three distinct varieties of adaptive tests can be distinguished (and are described below). But all varieties consist of two basic steps: question selection and score estimation. Both are repeated each time a question (or collection of questions) is presented and answered. The first step determines the most appropriate question (or collection of questions) to administer given what is currently known about the student's performance level. Selection is from a *pool,* which contains more questions than any single student is asked.

The second step uses the response(s) to the question(s) previously answered to refine the student's score or performance estimate. This allows the questions asked next to be more appropriate still.

This cycle continues until either a specified number of questions have been administered or some measure of score precision is reached. The process is represented schematically by Figure 1.

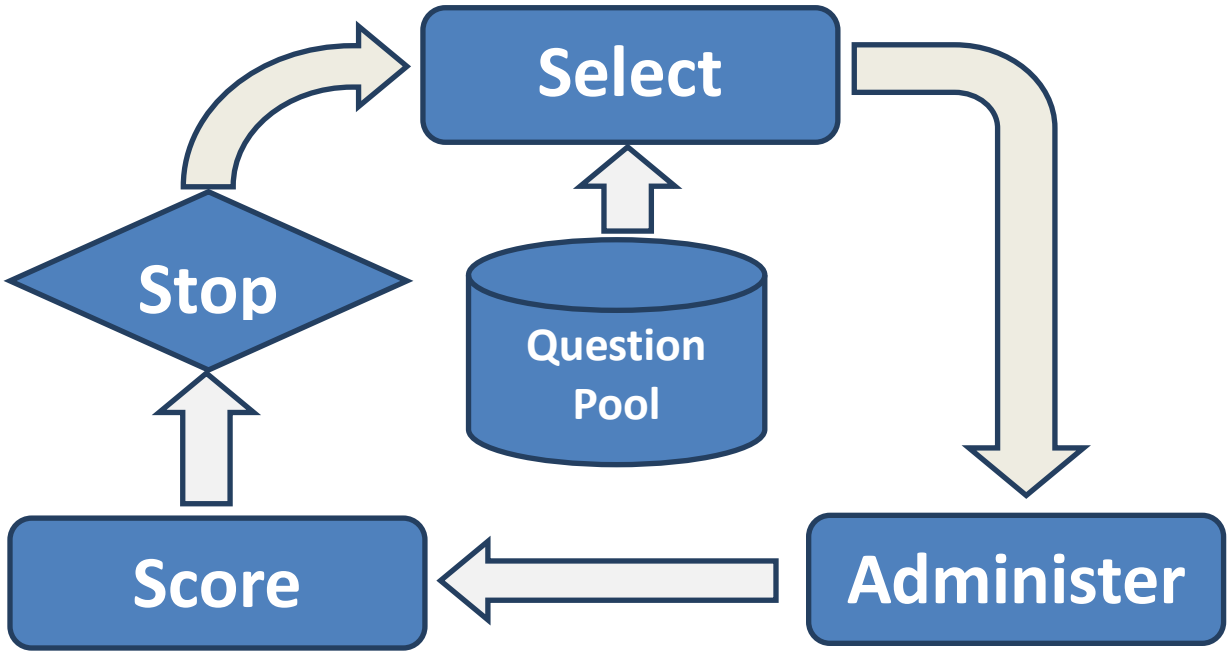

Figure 1: The adaptive testing cycle

Adaptive tests are designed to maximize *measurement efficiency*, or the precision of test scores in relation to test length. This means an adaptive test can either save time by being shorter than a conventional test of equal precision, or improve score quality by being more precise than a conventional test of equal length. The students with the most to gain are those at either the high or low extremes of the performance continuum. They are usually poorly served by conventional tests, which are generally designed to best fit the average student.

Researchers have developed and proposed numerous procedures for implementing each of the basic tasks needed to select and score an adaptive test. Methods have proliferated largely because none is ideal for all testing situations and circumstances. What works best instead depends on the unique characteristics of a given testing program. Test content, question formats, student characteristics, and even the subjective values of the test's developers and score users are all relevant considerations. The process of deciding among the various possibilities and choosing those best suited for a particular testing program starts by asking questions intended to determine exactly how a particular CAT system operates.[1]

[1] See Davey (2011), Davey & Pitoniak (2006), Parshall, Spray, Kalohn, & Davey (2002) or Wainer et al. (2000) for accessible and detailed accounts of how adaptive tests function.

QUESTION POOLS

As depicted in Figure 1, the question pool sits at the center of all adaptive test administration systems. The pool is an organized collection of questions, only some of which will be presented to any individual student during his/her test. The important properties of a question pool include size, composition, structure, and the nature of the item response model on which the questions are *calibrated*.

Pool size is conveniently quantified by multiples of test length. For example, if an adaptive test administers 30 questions to each student, then a 300-question pool contains 10 test lengths. In practice, the best size for a pool depends on a variety of factors. Pool size can be constrained or even dictated by circumstances (e.g., the number of questions available for use). At the same time, pool size determines many of the observed characteristics of a test (e.g., efficiency and security, the latter measured by the extent to which the tests administered to different students share questions in common).

Q1. How many questions do the pools contain and how were these numbers determined?

Strong answer: Pool size was determined by weighing the benefits of larger pools (greater efficiency and better security) against the practical and financial costs of developing and pretesting larger numbers of questions.

Weak answer: Pool size was dictated or limited by the number of items available for use, rather than by a rational process that weighed psychometric and practical considerations.

The *composition* of a pool is determined by the sort of questions it contains. The format, content, and performance characteristics of the questions all contribute to pool composition. As described below, the question selection procedure periodically enters the pool with a very specific shopping list in hand. An adaptive test works most effectively when the questions on that list are consistently found to be in stock.

Q2. Are there specifications (or blueprints) that dictate the composition of the question pool? If so, how were they determined?

Strong answer: Specifications were determined by simulating assembly of tests from question pools that varied in composition. The quality of these tests (both in terms of the statistical characteristics of the scores computed and the conformance of tests administered to content requirements) was then evaluated to determine a pool composition that allowed appropriate tests to be consistently produced.

Weak answer: Pool composition was dictated or limited by the items that were available for use. Little or no attempt was made to demonstrate that a pool so constituted could effectively and consistently support assembly of adaptive tests of appropriate quality.

The *structure* of the pool pertains to whether and how the individual questions are connected to one another. Questions can reside in the pool—and be selected and presented—as discrete, solitary objects. Alternatively, they might be organized into prearranged bundles, which are selected and presented as a unit. A good example is a reading passage to which several questions are attached. How the structure of the pool affects the way an adaptive test operates will be explored more thoroughly in the next section.

Finally, all questions in a pool must have been fit or *calibrated* under an item response theory (IRT) model. Adaptive tests rely on IRT operating behind the scenes to perform two key functions. The first is to produce scores that are comparable across students who have taken entirely different tests. The second is to identify those questions that are most appropriate to administer to a particular student given his/her demonstrated level of performance earlier in the test.

A variety of IRT models can be employed, with the differences between them subtle but potentially important.[2] However, all models require that substantial samples of response data be available for each question *before* it can be included in a pool. Data for the initial pool will therefore need to be collected and calibrated prior to the testing program becoming operational. A key assumption is that questions will perform the same way once operational as they did in the calibration sample. To the extent that they do not, the quality of reported scores can be compromised.

[2] See Baker (2001) or Hambleton, Swaminathan, & Rogers (1991) for good, accessible introductions to IRT.

Q3. When, how, and from whom were the data needed for IRT calibration collected?

Strong answer: Calibration data were collected (a) recently, (b) on computer (ideally using the same administration software as will be used operationally), and (c) from large, motivated, and representative samples of students.

Weak answers: Calibration data were (a) collected from previous, paper test forms (Will questions perform the same now as they did when those paper forms were administered? Will students perform identically on paper as they will on computer?), or (b) collected from small samples of potentially unmotivated students.

Q4. Are there plans to periodically refresh or replace the question pools over time? If so, how will the calibration data be collected?

Strong answer: New questions will be developed and routinely field-tested alongside or within operational adaptive tests, thus ensuring motivated, representative calibration samples.

Weak answers: (a) Pools will not be periodically refreshed with newly developed and calibrated questions. (b) Additional questions (and calibration data) will be harvested from other paper forms. (c) Newly developed question will be field-tested only on small, potentially unmotivated or unrepresentative student samples.

Q5. What item response theory model(s) are employed and why were they selected?

Strong answer: IRT models were selected following a principled, empirically based process. Evidence (e.g., goodness-of-fit measures) was provided to demonstrate that the selected models were appropriate for the students and test questions.

Weak answer: Models used were dictated by limitations in either the CAT delivery system or the logistics of data collection. Little or no evidence is provided that they were appropriate for the students and test questions modeled.

QUESTION SELECTION

Most CAT systems select questions for presentation in order to best measure the student being tested, subject to certain rules or restrictions. It's in the nature of these rules and in competing definitions of "best" where we find most of the differences in adaptive testing systems. These differences can be grouped into three major categories. The first concerns whether questions are pulled from the pool individually or in prearranged sets or bundles. In the "single question" case, the select-administer-score cycle depicted in Figure 1 is repeated as many times as the test is long: once for each question

administered. In the "bundled" case, the cycle is repeated fewer times, with multiple questions selected on each loop.

The extreme version of bundling is sometimes called *multistage testing* (MST). Although an MST conceptually operates in accordance with Figure 1, it is more clearly described by the diagram in Figure 2.

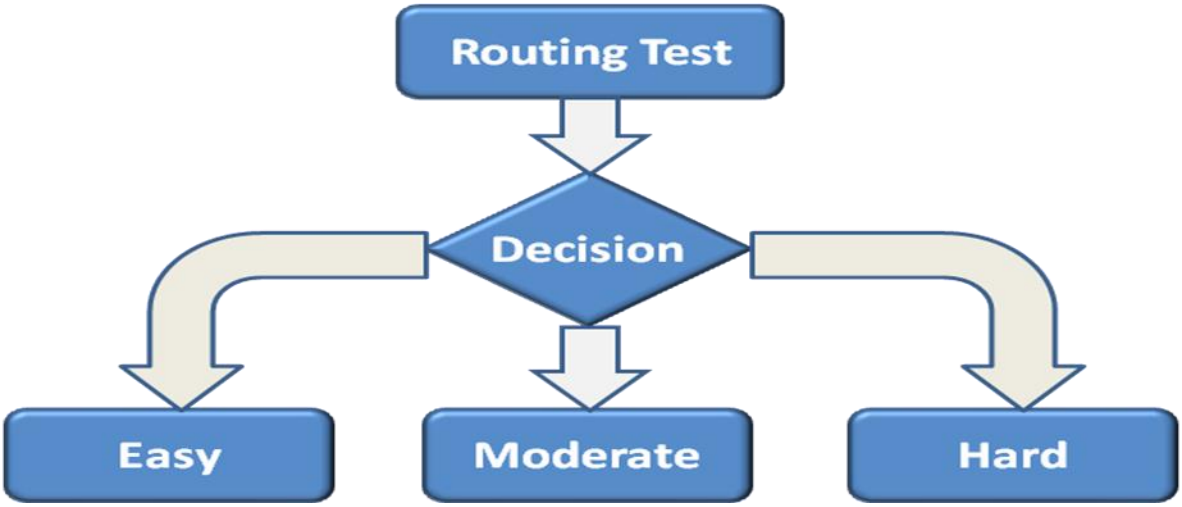

Figure 2: A two-stage MST

An MST starts by administering to each student a moderately difficult bundle of items, often called a *routing test*. Depending on their performance across this whole bundle, students are assigned to one of several second-stage tests that can differ widely in difficulty. Although the example depicted above ends at this point, further decisions and stages can follow. However, gains in efficiency generally rapidly diminish beyond a two- or three-stage design.

Q6. Are adaptive tests delivered according to the single-question or multistage (MST) design? Why was the design implemented chosen?

Strong answer: The test design was chosen after careful consideration of the strengths and weaknesses of each approach relative to the requirements of the testing program.

Weak answer: The approach implemented was dictated or limited by the capabilities of the test delivery system. Alternative test designs were not considered and no evidence is provided to demonstrate that the design employed best meets program requirements.

Whether selected individually or in bundles, questions are chosen to help a test achieve one or more measurement goals or objectives. The most common of these is to maximize the precision with which each student is measured. However, one drawback with this goal is that it can lead to widely varying levels of precision across students. Some students are simply easier to measure precisely than others. The extent of the disparity can suggest an alternative goal, which is to measure each student to some specified level of precision. A third possibility is to prioritize accurate classification of students into discrete categories (e.g., "Basic," "Proficient," "Advanced"). Interestingly, doing so can lead to very different tests than those delivered to maximize score precision.

Q7. What are the goals of the question-selection mechanism? Why were those goals chosen?

Strong answer: Measurement goals were chosen to best meet the requirements of the testing program. Evidence is provided that a variety of goals were considered and evaluated, and the goals that were judged to be best given program requirements were then chosen.

Weak answer: The goal or goals implemented were dictated or limited by the capabilities of the test delivery system. Test developers were either unaware of alternatives or present no evidence of having considered and evaluated them.

Q8. How, technically, are measurement-precision goals incorporated into the selection mechanism? Does the methodology accommodate questions scored both dichotomously (right/wrong) and polytomously (partial credit)?

Strong answer: A detailed description of the rules and processes the selection algorithm employs to choose questions was readily provided, ideally with illustrative examples. Performance of the methods employed was supported by research-based evidence.

Weak answer: No description was available, or a vague or incomplete description was provided. No research was reported or cited demonstrating performance of the selection mechanism.

Selection can (and in most cases should) be subject to considerations beyond measurement precision. The first consideration addresses the substance or content that the test is intended to measure. Content requirements or specifications can be characterized through blueprints like those that have driven conventional, paper-form assembly for decades. These blueprints specify the number or proportion of questions from various content domains or of various types or formats that a proper test is to contain. A number of mathematical algorithms can be used to force selection of tests that conform to the specifications while simultaneously pursuing one or more of the measurement goals outlined above.

Q9. Are content specifications imposed on question selection? If so, how were they developed? Are they as detailed as blueprints governing the assembly of conventional, paper forms?

Strong answer: Content specifications are in place as appropriate and were developed using the same process and at the same level of detail as blueprints for conventional test assembly.

Weak answer: No or minimal content specifications are imposed.

Q10. If required, how does the selection mechanism impose content specifications?

Strong answer: Content specifications are imposed on selection via some sort of "constrained optimization" algorithm. Technical information on this algorithm is made available as required.

Weak answer: Minimal specifications are imposed by a simple, deterministic process (e.g., by selecting the first five items from content "A," the next five from "B," etc.).

Q11. If required, how is selection of constructed-response (CR) questions handled?

Strong answer: A clear, detailed description of when and how CR questions are selected is provided.

Weak answer: No description is available or a vague and incomplete description is provided.

Under the single-question CAT design, the selection algorithm needs to be quite clever. The task of assembling an adaptive test is more challenging than that of assembling a conventional test because it needs to take place on the fly without the possibility of revision. This means that the selection mechanism (and the content specifications driving it) must be trusted implicitly to create a test form that can be administered without review. This is much less the case under the bundled or MST approach, in which the number of routes or distinct combinations of questions is often small enough to allow review of each prior to administration.

It is possible for a test to meet all assembly specifications and still be judged upon review as less than ideal (or even unacceptable). What might be termed *enemy questions* are often the culprits. Two questions are considered as enemies if they are too similar to one another, cue one another, or interact with one another in any of a number of ways that would lead test developers to conclude that they should not appear together on the same test form. Because of the varied nature of these interactions, identifying

and tracking “enemies" can be an onerous task, particularly as the question bank grows large. Once identified, the selection mechanism can include logic that prevents enemy items from appearing together.

Q12. Does the selection mechanism guarantee that every test delivered meets all required content specifications? If not, what proportion of tests is expected to be nonconforming?

Strongest answer: Yes, specifications are guaranteed to be met for all tests delivered. The selection mechanism has been proven by theoretical and practical research to ensure conformance of all tests, provided the question pool is properly comprised.

Strong answer: Although the selection mechanism does not guarantee conformance of all tests, empirical research shows that the vast majority of tests, in fact, conform.

Weak answer: No information is available regarding the proportion of conforming tests.

Q13. Does the test design permit prior review of all possible test forms that can be administered?

Strongest answer: Yes, all possible forms can be reviewed.

Strong answer: No review of actual tests is possible, but sample tests produced by the selection mechanism will be available for review prior to implementation to ensure that (a) content specifications are detailed enough, and (b) the selection mechanism sufficiently meets those constraints.

Weak answer: No review is possible and sample tests are not available.

Q14. Are “enemy questions” identified during the development process? If so, how does the selection mechanism prevent their appearing in the same form?

Strong answers: (a) Enemy questions are identified during the question development process and the selection mechanism effectively prevents enemies from appearing in the same test. (b) The delivery model (e.g., MST) allows prospective tests to be reviewed and revised prior to administration.

Weak answers: (a) No process exists for identifying enemy questions. (b) The selection mechanism is unable to prevent enemies from appearing in the same test.

Another typical requirement imposed on question selection involves the security of the question pool. Naturally, security concerns are most pronounced when important stakes are attached to the use of test scores. Adaptive tests are often administered from the same question pools for an extended period of time. Use of the same questions over time increases the chances of their becoming known to students prior to their testing. However, because the pool contains more questions than any student will see and

because the particular combination of questions a given student sees is unpredictable, this situation may not be viewed as excessively risky. It is at least much more secure than administering the same fixed form repeatedly.

The security of an adaptive test can be roughly gauged by the extent to which the tests administered to different students are observed to share questions or *overlap*. Overlap is determined both by the size of the pool and by the extent to which questions are used (or *exposed*) at balanced or equal rates. Unfortunately, if questions are selected based only on measurement and content properties, exposure rates may be anything but balanced. This realization gave rise to what are usually termed *exposure control* procedures. These procedures work to limit the exposure of frequently selected questions and force use of less commonly selected questions. Use of exposure control is much more common under the single-question CAT design than under MST.

Q15. Does the selection mechanism make use of an exposure control procedure? If so, how does it work? If not, was it explicitly regarded as unnecessary?

Strong answers: (a) Yes, an effective exposure control procedure is employed. A detailed description of that procedure is made available as required. (b) A statement (and compelling rationale) that exposure control was not employed because it was judged unnecessary given the test's purpose is also acceptable.

Weak answers: (a) No exposure control was employed (without justification for it not being used). (b) No description or uncertainty of what the test delivery system does is offered.

Q16. What is the expected level of overlap in tests administered to different students? What about tests administered to the same student on different occasions?

Strong answer: Expected overlap rates are provided along with a compelling justification that they are appropriate given the test's purpose and stakes.

Weak answer: Overlap rates cannot be determined or are otherwise unavailable.

QUESTION ADMINISTRATION

The psychometric properties of test scores can be affected by the quality and characteristics of the *interface*, or the software that presents questions and collects responses. Ideally, the interface should have a neutral impact on students, neither promoting nor impeding performance with regard to the constructs being measured. This requires the software designers to tread a fine line between an interface that is simple for students to use but which lacks important functionality, and one that supplies so many

functions that some students find it difficult or confusing to use (Schneiderman & Plaisant, 2009). Erring in either direction risks introducing noise or error to the measurement process, with scores revealing as much about a student's ability (and experience) with computers as they do about knowledge of the intended construct.

Q17. Was the test interface designed in accordance with accepted *human-computer interaction* (HCI) principles? Was the design confirmed through formal usability studies?

Strong answer: Yes, accepted HCI principles were employed in the design of the interface and that design was subsequently refined or confirmed through usability studies conducted on students similar to those who will be operationally tested. Ideally, the results of these studies are available for review.

Weak answer: The interface design appears awkward and potentially confusing. No evidence of usability studies can be supplied, or such studies that are available were conducted on students very different than those who are to be tested.

Q18. What tools and capabilities are students provided with during their test? Is each of these necessary to permit or facilitate measurement of the intended construct? Are any capabilities that could facilitate measurement absent?

Strong answer: A complete description of interface tools and capabilities is supplied. Each capability facilitates measurement by having a clear and well-reasoned relationship to the construct being assessed. The interface appropriately varies across content areas and grade levels, depending on the requirements of the content being presented and the age and abilities of the students being measured.

Weak answer: No or incomplete descriptions of capabilities and their relationship with constructs are provided. The interface appears to lack capabilities important for proper measurement or provides those capabilities in clumsy and confusing ways. Little customization of the interface to best suit different content areas or grade levels is evident.

Q19. Are students provided with sufficient instruction and practice prior to their test?

Strong answer: Yes. Instructions are clear, concise, and presented appropriately for the students to be tested. Ideally, evidence from usability studies showing that the presentation is effective is also provided.

Weak answer: Instructions are incomplete, confusingly presented, or otherwise likely to be inappropriate or ineffective for the students to be tested. No evidence from usability studies is available regarding the instructions' effectiveness.

Q20. Are test-administration hardware and software standardized across (unaccommodated) students (e.g., interface, screen size and resolution, input devices, etc.)? If not, was evidence collected to show that variation does not impact scores?

Strong answers: (a) Yes, conditions are standardized. (b) Well-designed research was conducted showing that variation does not matter.

Weak answer: Conditions are not standardized and the impact of this on student performance is unknown.

Research has been conducted on the use of technology to make computer-based testing accessible to students with various physical and cognitive disabilities (see Johnstone, Altman, & Thurlow [2006] for a review). The term *universal design* is often applied to tests and test-delivery systems that were designed from the outset to serve the widest possible range of students. The principles of universal design are codified in the *accessible portable item profile* (APIP) standard, which was developed to promote adoption of accessibility technology by test development and delivery vendors.

Q21. Do the questions and the test delivery platform support the *accessible portable item profile* (APIP) standard? If not, what accommodations are offered at the question and test level?

Strong answers: (a) Yes, APIP is supported. (b) A detailed description of a wide array of accommodations is made available.

Weak answer: APIP is not supported and only a vague description of limited accommodation capabilities is available.

SCORING AND SCORE PROPERTIES

The most common method for scoring adaptive tests is through IRT *proficiency* estimates. These scores have the advantage of being comparable even across students who take very different tests. However, they are expressed on a scale that is usually considered inconvenient for reporting purposes. (Proficiency estimates typically range from about -5 to +5, with a mean or average around 0.) As such, most adaptive tests distinguish between the interim or *provisional* scores produced while the test is in process and which drive selection of future questions, and the *final* scores that are ultimately reported. Final scores are usually transformations of proficiency estimates to a more convenient or accessible scale.

Both provisional and final scores are distinguished by three important properties: bias, precision, and substantive meaning. Bias is the extent to which a score does or does not eventually converge on the true proficiency value if the test were allowed to continue administering questions indefinitely. Precision refers to the extent to which a score is contaminated by random measurement error. Generally speaking, precise scores are less variable across repeated testing of the same examinee than imprecise scores are. Both bias and precision are statistical properties and so can be quantified in a variety of standard ways (e.g., test reliability, standard errors of measurement, test information functions, etc.).

Q22. How are provisional and reported scores computed?

Strong answer: A detailed, technically defensible description of the methodology is provided.

Weak answer: Only a vague or technically questionable description is provided.

Q23. What are the statistical characteristics of reported scores? How are these expressed? What assumptions are implicit in those expressions?

Strong answer: Statistical characteristics of reported scores are described in detail, along with the assumptions inherent in computing them.

Weak answer: Vague or technically questionable descriptions are provided.

Q24. How do the statistical characteristics of adaptive test scores compare to those of conventional tests used for comparable purposes with comparable student populations? How does measurement of those characteristics differ between adaptive tests and conventional tests?

Strong and weak answers are the same as for Q23.

Q25. To what extent does measurement quality differ or vary across students?

Strong answer: Traditional measures of measurement quality (such as test reliability or test information functions) are averaged or aggregated across examinees, and therefore hide the fact that some students are measured better than others. A strong answer would recognize this by providing an estimate of the variability of disaggregated or student-specific measurement characteristics, both across particular students and across score ranges.

Weak answer is the same as the answer to Q23.

Q26. How was the length of the adaptive test determined?

Strong answer: Length was determined by a principled, research-based process that weighed the quality of scores required by the stakes attached to those scores against practical considerations of testing time and question-development expense.

Weak answer: Length was determined arbitrarily or entirely driven by practical constraints.

Q27. If a time limit is imposed on the test, how was it determined?

Strong answer: Appropriate time limits were determined by pilot testing or substantial previous experience with the questions, delivery system, and students involved.

Weak answer: Time limits were determined arbitrarily or driven entirely by practical constraints.

The substantive meaning of a score is harder to characterize than its statistical properties. Meaning is attached to a score through the test specifications or blueprint, which indicate the numbers of questions of each content classification or type that comprise a proper test. Important content areas or question types exert more influence on scores by being more heavily represented in each test form. This works reasonably well when all questions contribute equally to total scores, as is the case with the number-right scores typically computed for conventional tests.

However, questions usually do not contribute equally to IRT proficiency estimates. Instead, questions influence proficiency estimates in proportion to the strength with which they measure that proficiency. Examples can readily be devised of content areas or question types exerting an influence on scoring far out of proportion to their representation in the test specifications. Technically, the amount of influence a question or content area exerts is measured by the IRT *information function.* These functions generally vary substantially across the proficiency range, complicating matters further. It is therefore not unusual for lower scores to be driven primarily by one content area while higher scores are more heavily influenced by another. See Davey and Pitoniak (2006) for a more complete description of this phenomenon.

Q28. How is IRT information distributed across the content areas and question types that comprise the test blueprint? Are the proportions of information contributed from each area in accordance with substantive requirements?

Strong answer: Information functions summed across relevant content areas and question types are provided and found to be in line with requirements.

Weak answer: Information functions are not available for specific content areas or are out of line with requirements.

Adaptive tests often need to be treated as comparable with preceding or continuing conventional paper-based tests. In the best case, the adaptive test entirely replaces a conventional test but needs to maintain continuity of performance trends, growth measures, or proficiency standards. In the worst case, the adaptive test must be used interchangeably with paper tests that continue to be administered to some students. In either case, comparability of the CAT and paper-test scores should be evaluated by research studies. However, the evidence for comparability must be much stronger in the latter case than in the former.

Studies evaluating the comparability of CAT and conventional testing programs can differ enormously in quality. Data can be collected under strong experimental designs (single or equivalent group) or weaker quasi-experimental designs (historical comparison, matched groups, etc.). Student samples can be large, representative, and motivated—or enjoy none of these attributes. Analyses can be well-formulated and properly analyzed, or vaguely defined and incomplete.

Q29. If required, is strong evidence of comparability with prior or parallel conventional tests provided?

Strong answer: A well-designed study was conducted and sufficient data samples were collected and properly analyzed, with results showing high comparability between CAT and conventional test performance.

Weak answer: The comparability study design was weak, small data samples were collected, or results proved inconclusive.

CONCLUSION

The questions posed above illustrate some of the many ways that adaptive tests can differ. Many of these differences are technical, subtle, and invisible to most students and score users. But these differences can have significant impact on the quality, meaning, and usefulness of the scores produced. Adaptive testing is a complex tool best wielded by those experienced and skilled in its use. Whether and how these questions are answered can help in judging whether the necessary expertise has in fact been exercised.

The focus here on psychometric issues has left unexplored an even larger array of questions concerning test delivery (e.g., Does the delivery system require that an Internet connection be maintained throughout testing?), question content (e.g., Is appropriate use made of innovative, computer-delivered question types?), and the burgeoning field of artificial-intelligence scoring of constructed-response questions (e.g., Can a certain writing sample be scored by a computer or are human raters required?). To fully understand the capabilities and limitations of any adaptive testing system requires consideration of these matters as well.

REFERENCES

Baker, F.B. (2001). *The basics of item response theory* (2nd ed.). College Park, MD: ERIC Clearinghouse on Assessment and Evaluation, University of Maryland.

Davey, T. (2011). *Practical considerations in computer-based testing* (Educational Testing Service Research Rep. No. CBT-2011). Princeton, NJ: Educational Testing Service.

Davey, T., & Pitoniak, M.J. (2006). Designing computerized adaptive tests. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development.* Mahwah, NJ: Lawrence Erlbaum.

Hambleton, R., Swaminathan, H., & Rogers, J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage Publications.

Johnstone, C.J., Altman, J., & Thurlow, M. (2006). *A state guide to the development of universally designed assessments*. Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Parshall, C., Spray, J., Kalohn, J., & Davey, T. (2002). *Practical considerations in computerized testing*. New York: Springer-Verlag.

Schneiderman, B., & Plaisant, C. (2009). *Designing the user interface: Strategies for effective human-computer interaction* (5th ed.). Reading, MA: Addison-Wesley.

Wainer, H., Dorans, N., Flaugher, R., Mislevy, R., Green, B., Steinberg, L. & Thissen, D. (2000). *Computerized adaptive testing: A primer* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

APPENDIX: SUMMARY WORKBOOK

QUESTION POOLS		
Question	**Strong & Weak Answers**	**Notes**
Q1. How many questions do the pools contain and how were these numbers determined?	**Strong answer:** Pool size was determined by weighing the benefits of larger pools (greater efficiency and better security) against the practical and financial costs of developing and pretesting larger numbers of questions. **Weak answer:** Pool size was dictated or limited by the number of items available for use, rather than by a rational process that weighed psychometric and practical considerations.	
Q2. Are there specifications (or blueprints) that dictate the composition of the question pool? If so, how were they determined?	**Strong answer**: Specifications were determined by simulating assembly of tests from question pools that varied in composition. The quality of these tests (both in terms of the statistical characteristics of the scores computed and the conformance of tests administered to content requirements) was then evaluated to determine a pool composition that allowed appropriate tests to be consistently produced. **Weak answer:** Pool composition was dictated or limited by the items that were available for use. Little or no attempt was made to demonstrate that a pool so constituted could effectively and consistently support assembly of adaptive tests of appropriate quality.	

QUESTION POOLS (CONT.)

Question	Strong & Weak Answers	Notes
Q3. When, how, and from whom were the data needed for IRT calibration collected?	**Strong answer:** Calibration data were collected (a) recently, (b) on computer (ideally using the same administration software as will be used operationally), and (c) from large, motivated, and representative samples of students. **Weak answers**: Calibration data were (a) collected from previous, paper test forms (Will questions perform the same now as they did when those paper forms were administered? Will students perform identically on paper as they will on computer?), or (b) collected from small samples of potentially unmotivated students.	
Q4. Are there plans to periodically refresh or replace the question pools over time? If so, how will the calibration data be collected?	**Strong answer:** New questions will be developed and routinely field-tested alongside or within operational adaptive tests, thus ensuring motivated, representative calibration samples. **Weak answers:** (a) Pools will not be periodically refreshed with newly developed and calibrated questions. (b) Additional questions (and calibration data) will be harvested from other paper forms. (c) Newly developed question will be field-tested only on small, potentially unmotivated or unrepresentative student samples.	
Q5. What item response theory model(s) are employed and why were they selected?	**Strong answer:** IRT models were selected following a principled, empirically based process. Evidence (e.g., goodness-of-fit measures) was provided to demonstrate that the selected models were appropriate for the students and test questions. **Weak answer:** Models used were dictated by limitations in either the CAT delivery system or the logistics of data collection. Little or no evidence is provided that they were appropriate for the students and test questions modeled.	

QUESTION SELECTION		
Question	**Strong & Weak Answers**	**Notes**
Q6. Are adaptive tests delivered according to the single-question or multistage (MST) design? Why was the design implemented chosen?	**Strong answer:** The test design was chosen after careful consideration of the strengths and weaknesses of each approach relative to the requirements of the testing program. **Weak answer:** The approach implemented was dictated or limited by the capabilities of the test delivery system. Alternative test designs were not considered and no evidence is provided to demonstrate that the design employed best meets program requirements.	
Q7. What are the goals of the question-selection mechanism? Why were those goals chosen?	**Strong answer:** Measurement goals were chosen to best meet the requirements of the testing program. Evidence is provided that a variety of goals were considered and evaluated, and the goals employed that were judged to be best given program requirements were then chosen. **Weak answer:** The goal or goals implemented were dictated or limited by the capabilities of the test delivery system. Test developers were either unaware of alternatives or present no evidence of having considered and evaluated them.	
Q8. How, technically, are measurement-precision goals incorporated into the selection mechanism? Does the methodology accommodate questions scored both dichotomously (right/wrong) and polytomously (partial credit)?	**Strong answer:** A detailed description of the rules and processes the selection algorithm employs to choose questions was readily provided, ideally with illustrative examples. Performance of the methods employed was supported by research-based evidence. **Weak answer:** No description was available, or a vague or incomplete description was provided. No research was reported or cited demonstrating performance of the selection mechanism.	

QUESTION SELECTION (CONT.)		
Question	**Strong & Weak Answers**	**Notes**
Q9. Are content specifications imposed on question selection? If so, how were they developed? Are they as detailed as blueprints governing the assembly of conventional, paper forms?	**Strong answer:** Content specifications are in place as appropriate and were developed using the same process and at the same level of detail as blueprints for conventional test assembly. **Weak answer:** No or minimal content specifications are imposed.	
Q10. If required, how does the selection mechanism impose content specifications?	**Strong answer:** Content specifications are imposed on selection via some sort of constrained optimization algorithm. Technical information on this algorithm is made available as required. **Weak answer:** Minimal specifications are imposed by a simple, deterministic process (e.g., by selecting the first five items from content "A," the next five from "B," etc.).	
Q11. If required, how is selection of constructed-response (CR) questions handled?	**Strong answer:** A clear, detailed description of when and how CR questions are selected is provided. **Weak answer**: No description is available or a vague and incomplete description is provided.	

QUESTION SELECTION (CONT.)

Question	Strong & Weak Answers	Notes
Q12. Does the selection mechanism guarantee that every test delivered meets all required content specifications? If not, what proportion of tests is expected to be nonconforming?	**Strongest answer:** Yes, specifications are guaranteed to be met for all tests delivered. The selection mechanism has been proven by theoretical and practical research to ensure conformance of all tests, provided the question pool is properly comprised. **Strong answer:** Although the selection mechanism does not guarantee conformance of all tests, empirical research shows that the vast majority of tests, in fact, conform. **Weak answer**: No information is available regarding the proportion of conforming tests.	
Q13. Does the test design permit prior review of all possible test forms that can be administered?	**Strongest answer:** Yes, all possible forms can be reviewed. **Strong answer:** No review of actual tests is possible, but sample tests produced by the selection mechanism will be available for review prior to implementation to ensure that (a) content specifications are detailed enough, and (b) the selection mechanism sufficiently meets those constraints. **Weak answer:** No review is possible and sample tests are not available.	
Q14. Are "enemy questions" identified during the development process? If so, how does the selection mechanism prevent their appearing in the same form?	**Strong answers:** (a) Enemy questions are identified during the question development process and the selection mechanism effectively prevents enemies from appearing in the same test. (b) The delivery model (i.e., MST) allows prospective tests to be reviewed and revised prior to administration. **Weak answers:** (a) No process exists for identifying enemy questions. (b) The selection mechanism is unable to prevent enemies from appearing in the same test.	

QUESTION SELECTION (CONT.)		
Question	**Strong & Weak Answers**	**Notes**
Q15. Does the selection mechanism make use of an exposure control procedure? If so, how does it work? If not, was it explicitly regarded as unnecessary?	**Strong answers:** (a) Yes, an effective exposure control procedure is employed. A detailed description of that procedure is made available as required. (b) A statement (and compelling rationale) that exposure control was not employed because it was judged unnecessary given the test's purpose is also acceptable. **Weak answers:** (a) No exposure control was employed (without justification for it not being used). (b) No description or uncertainty of what the test delivery system does is offered.	
Q16. What is the expected level of overlap in tests administered to different students? What about tests administered to the same student on different occasions?	**Strong answer:** Expected overlap rates are provided along with a compelling justification that they are appropriate given the test's purpose and stakes. **Weak answer:** Overlap rates cannot be determined or are otherwise unavailable.	

QUESTION ADMINISTRATION		
Question	**Strong & Weak Answers**	**Notes**
Q17. Was the test interface designed in accordance with accepted *human-computer interaction* (HCI) principles? Was the design confirmed through formal usability studies?	**Strong answer:** Yes, accepted HCI principles were employed in the design of the interface and that design was subsequently refined or confirmed through usability studies conduct on students similar to those who will be operationally tested. Ideally, the results of these studies are available for review. **Weak answer:** The interface design appears awkward and potentially confusing. No evidence of usability studies can be supplied, or such studies that are available were conducted on students very different than those who are to be tested.	
Q18. What tools and capabilities are students provided with during their test? Is each of these necessary to permit or facilitate measurement of the intended construct? Are any capabilities that could facilitate measurement absent?	**Strong answer:** A complete description of interface tools and capabilities is supplied. Each capability facilitates measurement by having a clear and well-reasoned relationship to the construct being assessed. The interface appropriately varies across content areas and grade levels, depending on the requirements of the content being presented and the age and abilities of the students being measured. **Weak answer:** No or incomplete descriptions of capabilities and their relationship with constructs are provided. The interface appears to lack capabilities important for proper measurement or provides those capabilities in clumsy and confusing ways. Little customization of the interface to best suit different content areas or grade levels is evident.	
Q19. Are students provided with sufficient instruction and practice prior to their test?	**Strong answer:** Yes. Instructions are clear, concise, and presented appropriately for the students to be tested. Ideally, evidence from usability studies showing that the presentation is effective is also provided. **Weak answer:** Instructions are incomplete, confusingly presented, or otherwise likely to be inappropriate or ineffective for the students to be tested. No evidence from usability studies is available regarding the instructions' effectiveness.	

QUESTION ADMINISTRATION (CONT.)		
Question	**Strong & Weak Answers**	**Notes**
Q20. Are test-administration hardware and software standardized across (unaccommodated) students (e.g., interface, screen size and resolution, input devices, etc.)? If not, was evidence collected to show that variation does not impact scores?	**Strong answers:** (a) Yes, conditions are standardized. (b) Well-designed research was conducted showing that variation does not matter. **Weak answer:** Conditions are not standardized and the impact of this on student performance is unknown.	
Q21. Do the questions and the test delivery platform support the *accessible portable item profile* (APIP) standard? If not, what accommodations are offered at the question and test level?	**Strong answers:** (a) Yes, APIP is supported. (b) A detailed description of a wide array of accommodations is made available. **Weak answer:** APIP is not supported and only a vague description of limited accommodation capabilities is available.	

SCORING AND SCORE PROPERTIES		
Question	**Strong & Weak Answers**	**Notes**
Q22. How are provisional and reported scores computed?	**Strong answer:** A detailed, technically defensible description of the methodology is provided. **Weak answer:** Only a vague or technically questionable description is provided.	
Q23. What are the statistical characteristics of reported scores? How are these expressed? What assumptions are implicit in those expressions?	**Strong answer:** Statistical characteristics of reported scores are described in detail, along with the assumptions inherent in computing them. **Weak answer:** Vague or technically questionable descriptions are provided.	
Q24. How do the statistical characteristics of adaptive test scores compare to those of conventional tests used for comparable purposes with comparable student populations? How does measurement of those characteristics differ between adaptive tests and conventional tests?	Strong and weak answers are the same as for Q23.	

SCORING AND SCORE PROPERTIES (CONT.)		
Question	**Strong & Weak Answers**	**Notes**
Q25. To what extent does measurement quality differ or vary across students?	**Strong answer:** Traditional measures of measurement quality (such as test reliability or test information functions) are averaged or aggregated across examinees, and therefore hide the fact that some students are measured better than others. A strong answer would recognize this by providing an estimate of the variability of disaggregated or student-specific measurement characteristics, both across particular students and across score ranges. **Weak answer** is the same as the answer to Q23.	
Q26. How was the length of the adaptive test determined?	**Strong answer**: Length was determined by a principled, research-based process that weighed the quality of scores required by the stakes attached to those scores against practical considerations of testing time and question-development expense. **Weak answer:** Length was determined arbitrarily or entirely driven by practical constraints.	
Q27. If a time limit is imposed on the test, how was it determined?	**Strong answer:** Appropriate time limits were determined by pilot testing or substantial previous experience with the questions, delivery system, and students involved. **Weak answer:** Time limits were determined arbitrarily or driven entirely by practical constraints.	

SCORING AND SCORE PROPERTIES (CONT.)		
Question	**Strong & Weak Answers**	**Notes**
Q28. How is IRT information distributed across the content areas and question types that comprise the test blueprint? Are the proportions of information contributed from each area in accordance with expectations?	**Strong answer:** Information functions summed across relevant content areas and question types are provided and found to be in line with requirements. **Weak answer:** Information functions are not available for specific content areas or are out of line with requirements.	
Q29. If required, is strong evidence of comparability with prior or parallel conventional tests provided?	**Strong answer:** A well-designed study was conducted and sufficient data samples were collected and properly analyzed, with results showing high comparability between CAT and conventional test performance. **Weak answer:** The comparability study design was weak, small data samples were collected, or results proved inconclusive.	