An NCME Instructional Module on Multistage Testing

Amy Hendrickson, *The College Board*

Multistage tests are those in which sets of items are administered adaptively and are scored as a unit. These tests have all of the advantages of adaptive testing, with more efficient and precise measurement across the proficiency scale as well as time savings, without many of the disadvantages of an item-level adaptive test. As a seemingly balanced compromise between linear paper-and-pencil and item-level adaptive tests, development and use of multistage tests is increasing. This module describes multistage tests, including two-stage and testlet-based tests, and discusses the relative advantages and disadvantages of multistage testing as well as considerations and steps in creating such tests.

Keywords: adaptive, multistage, testlet

Multistage tests are those in which preconstructed sets of items are administered adaptively and are scored as a unit. With multistage tests, adaptation occurs at the item set level. This results in fewer adaptation points than with item-level computerized adaptive tests (CATs), in which adaptation occurs after every item, but more adaptation points than in conventional paper-and-pencil linear tests, in which all examinees are administered all of the same questions. Multistage tests combine the advantages of both adaptive and linear test forms (Berger, 1994). As such, multistage tests, both two-stage and testlet-based, provide a balanced compromise between computerized adaptive tests and linear tests, which has led to their increasingly widespread research and use.

Amy Hendrickson is an Associate Psychometrician for The College Board, 1233 20th Street NW, Suite 600, Washington, DC 20036-2375; ahendrickson@collegeboard.org. Her interests are measurement, test equating, and vertical scaling.

Series Information

ITEMS is a series of units designed to facilitate instruction in educational measurement. These units are published by the National Council on Measurement in Education. This module may be photocopied without permission if reproduced in its entirety and used for instructional purposes. Dr. Deborah Harris has served as editor for this module. Information regarding the development of new ITEMS modules should be addressed to: Dr. Mark Gierl, Canada Research Chair in Educational Measurement and Director, Centre for Research in Applied Measurement and Evaluation, Department of Educational Psychology, 6-110 Education North, University of Alberta, Edmonton, Alberta, Canada T6G 2G5.

Item-Level Adaptive Tests

Adaptive testing grew out of a desire for more efficient and precise measurement of examinees across the entire proficiency distribution compared to that accomplished by linear tests, and it has been shown to be advantageous in this task (Lord, 1980; Wainer, 1990). For conventional linear tests, in which each examinee takes the same items in the same order, precision of measurement varies across the range of examinees' proficiencies. More specifically, the highest precision of a linear test often exists at a point representing mean performance of the intended measurement group. Thus, conventional linear tests measure examinees of average proficiency within the group quite well, but less precise measurements are made for those at the ends of the proficiency scale (Hambleton & Swaminathan, 1985; Lord, 1980; Weiss, 1974). Adaptive tests, on the other hand, focus measurement at an individual test-taker's proficiency level and thus can provide equally precise measurement for all examinees, often with fewer items than for a linear test.

Item-level adaptive tests present each new item based on the examinee's performance on all previous items and have been most widely researched. These tests have been shown to allow for shorter tests with equivalent or higher measurement precision when compared with conventional linear tests, especially for examinees with ability levels in the extremes of the distribution (Lord, 1974; Loyd, 1984; Wainer, Kaplan, & Lewis, 1992).

To illustrate this type of test, take for example a test of high school students' biology knowledge. With an item-level computerized adaptive test, individual items would be chosen based on a student's performance on all previous

items, administering easier or more difficult items to match the student's ability level. The algorithm may also incorporate content balancing so that items from different content areas (e.g., cell processes, ecosystems, genetics) would be included. Each student would potentially take a different set of items: those that are best targeted to their ability.

There are several potential problems with item-level adaptive tests, however. They include the potential for (1) violation of the item response theory (IRT) assumptions of local independence and unidimensionality, (2) lack of control over item ordering and the potential for context effects, (3) lack of control over nonstatistical properties including content balancing, (4) the need for item exposure control for test security, (5) lack of review opportunity for examinees, and (6) large data management and computer processing demands (Hambleton, Swaminathan, & Rogers, 1991; Vispoel, 1998; Wainer & Kiely, 1987; Wainer, Lewis, Kaplan, & Braswell, 1990, and Yen, 1993).

The assumptions of local item independence and unidimensionality state that, conditional on ability on the construct of interest, examinees' performances on each item are independent of their performance on all other items and that there is one underlying construct being measured by the test. Specifically, in the context of the biology exam, violation of local independence and unidimensionality may occur if performing well on the biology exam also requires high reading ability. Thus, biology knowledge as well as reading comprehension is assessed. Estimates of reliability may be inflated for item-level CATs if local item dependence is present. Discrimination parameters are inflated for items that are locally dependent, leading to inflated estimates of precision/reliability.

With item-level adaptive tests, item administration is based on each examinee's performance. Thus the test developer does not have control over the item order, such as easiest to hardest. This lack of control may also lead to context effects. A context effect may occur if for some students a question about photosynthesis follows a question about plants and the photosynthesis question is easier when this occurs.

To control the distribution of nonstatistical item characteristics, such as content balancing, on item-level CATs, each item's properties must be identified and the algorithms for item selection must incorporate all of these properties. Lack of content balancing could occur if a student's CAT includes too many items on the muscular system and not enough on the digestive system, according to the established table of specifications.

Another example of a potential problem with item-level adaptive tests is test security and item exposure control. Test security may be suspect if the starting point and adaptation within the biology exam occur without regard for controlling exposure of the items and if students are allowed to take the exam at different times. Thus, one student may take the exam and memorize the first few items that were administered and tell another student what these items were. The second student then would have an advantage on the test given this knowledge, if no attempt is made in the test algorithm to avoid using the same items repeatedly. Maintaining the security of the test and items requires item exposure control methods and large item pools.

On an item-level adaptive biology exam, students would most likely not be allowed to skip items or to go back and review and change their answers because of the nature of the adaptive administration and the potential for score inflation. For example, with an item-level CAT that allows for changing of answers at the end of the test, an examinee may intentionally answer all of the items incorrectly, thus "creating" a relatively easy test. During review, a student could then go back and answer all of the questions correctly. This may lead to an inflated ability estimate. See Wainer (1993), Wise (1996), and Kingsbury (1996) for discussion of this and other possible score inflation strategies.) Thus, because items are chosen based on previous responses, answer changes on item-level CATs are not permitted, despite examinees' strong desire for this opportunity (Vispoel, 1998; Vispoel, Hendrickson, & Bleiler, 2000).

Finally, item level CATs require lots of data management and computer processing demands for administering the tests and keeping track of each examinee's performance on each item. See Hambleton et al. (1991), Vispoel (1998), Wainer & Kiely (1987), Wainer et al. (1990), and Yen (1993) for more discussion of these issues.

Multistage Adaptive Tests

Test adaptation need not occur at the item level to improve precision and efficiency of measurement, however. Adaptation may occur between item sets, or testlets, based on cumulative performance on previous item sets, rather than between each item, as in a traditional CAT. These non-item level tests may help avoid the problems encountered with item-level adaptation. Non-item level tests contain fewer adaptation points when compared with item-level CATs, but more adaptation points than in paper-and-pencil linear tests. Such tests can be lumped under the heading of multistage tests and include two-stage and testlet-based tests (Betz & Weiss, 1973; Lord, 1980). Tests of this type have been investigated and/or put into operational use for such large-scale assessments as the Law School Admission Test (LSAT), the Test of English as a Foreign Language (TOEFL), the National Council of Architectural Registry Boards (NCARB), the National Assessment of Educational Progress (NAEP), and the U.S. Medical Licensure Examination (USMLE) (Bock & Zimowski, 1998; Luecht & Nungester, 1998; Schnipke & Reese, 1997; Wainer, 1995; Wainer et al., 1990; Wainer & Lukhele, 1997).

Figure 1 represents an example multistage test procedure. Multistage tests start with a first-stage, or routing, test that is generally short, is used for initial estimation, and may contain items with a broad range of difficulty values (very easy–very difficult) or may have a high concentration of items with difficulty located at the average or median proficiency of the intended group. An examinee's performance on this first-stage test is used to estimate their ability and to determine which second-stage test includes items with difficulty values that best match this ability and thus will lead to the most precise measurement (Betz & Weiss, 1973; Lord, 1980). This choice is made from several alternative second-stage, or measurement, tests containing items concentrated at a particular level of difficulty. Thus, these second-stage tests are used to differentiate among the abilities of individuals within a narrower range of proficiency than the routing test (Weiss, 1974). The test ends after completion of this second test if a two-stage test is desired, such that there is only one adaptation point. A two-stage test represents the simplest of the multistage adaptive test strategies and is actually two conventional linear tests, where the first test is scored

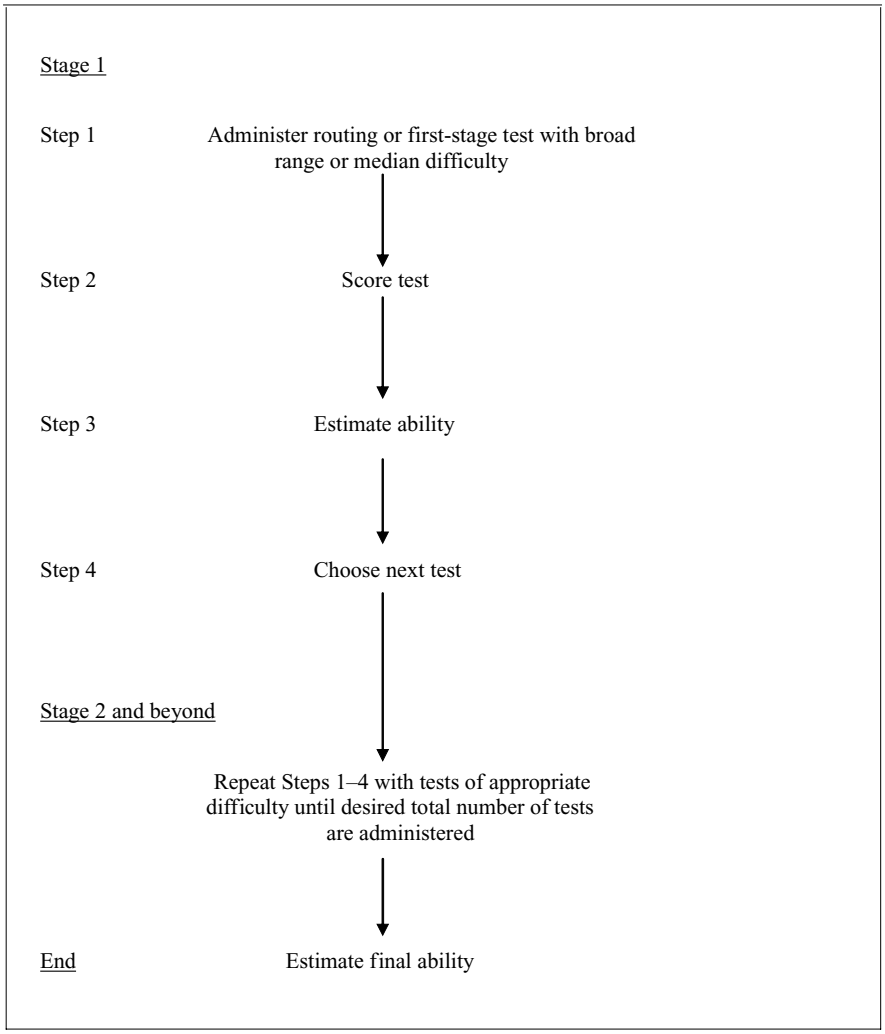

FIGURE 1. Example multistage test procedure.

before administering the second. The adaptive administration process continues in a multistage test, however, such that examinees are routed to tests with more narrowly focused difficulty at each stage based on their performance on previous stages.

Figure 2 contains an example of a five-stage test design with a routing test at the first stage and three to five item sets (T1-T16) for possible administration at Stages 2–5. This test design could be applied to the previously described biology test example, such that all students initially complete the routing test that includes biology items with a range of difficulty values. Then, based on their performance on this routing test, each student would be administered a second set of items during Stage 2 (T1, T2, or T3) containing biology items with difficulties that are a closer match to their ability. This adaptation would continue after completion of an item set within each of the first four stages of the five-stage test such that each examinee completes a total of five-item sets including the routing test.

Testlets

The item sets on multistage tests are often referred to as testlets. Wainer & Kiely (1987) described a testlet as a group of items which relate to a single content area, are constructed and analyzed as a unit, and are statistically independent of all other testlets and items. They suggested that testlets be created by content area specialists before the adaptive test is administered and be scored as polytomous items under a graded response model (see van der Linden & Hambleton (1997) for more information about polytomous IRT models).

A more general definition of testlets allows for the items to be stimulus-dependent, content-balanced, content-specific, or any other set of items, each of which is built to a particular range of ability (Sireci, Thissen, & Wainer, 1991; Thissen, Steinberg, & Mooney, 1989; Wainer, 1990; Wainer & Kiely, 1987; Wainer et al., 1990). A stimulus-dependent testlet may consist of a set of items that all relate to the same reading comprehension passage or to the same table or graph. For example, a testlet for the high school biology exam may consist of a diagram of a cell with five items concerning the function and description of each part of the cell. Another example is a testlet that includes a reading passage about genetic disorders with six items pertaining to the passage.

Other commonly used types of testlets are those that are content-balanced (within-testlet balanced) or content specific (between-testlet balanced). For example, a within-testlet balanced high school biology exam may consist of 5-item content-balanced testlets (such as T1-T16 in Figure 2), where each item in a testlet represents the content of one of

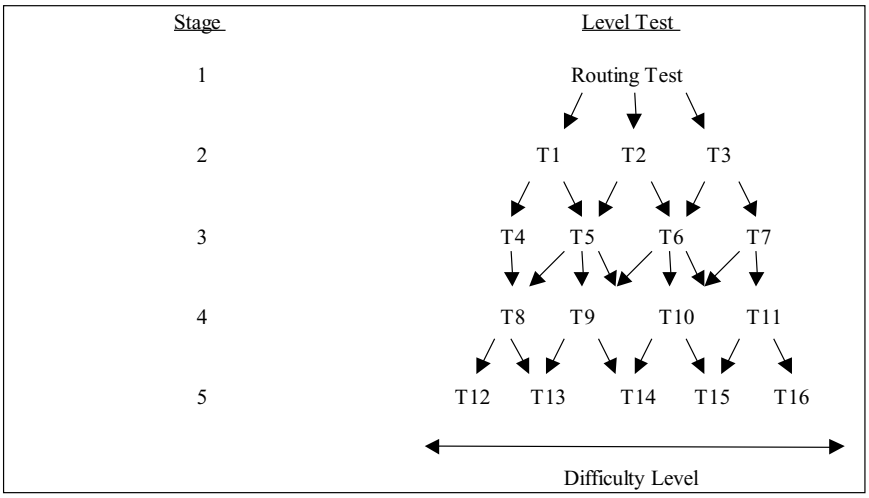

FIGURE 2. Example multistage test design.

the five areas to be covered (heredity, photosynthesis, evolution, etc.). Alternatively, a between-testlet balanced high school biology exam would be created from content-specific testlets within each stage, each containing five items that cover just one content area [e.g., cell processes at Stage 1 (T1-T3), ecosystems at Stage 2 (T4-T7), etc.], and through administration of one content-specific testlet in each of the five stages, the test is content-balanced.

In this module, testlets will refer to any set of items that are scored as a unit and are built to a particular difficulty range. Thus, with testlet-based multistage tests, the stages consist of one or more possible testlets and the adaptation decision between stages determines which testlet (of a certain difficulty range) an examinee will encounter next.

What are the relative advantages of a multistage test compared to a paper-and-pencil test?

There are several advantages of multistage tests when compared with linear tests. First, multistage tests allow for more efficient and precise measurement across the proficiency scale compared to linear tests, including at the extremes of the proficiency range. Thus, these adaptive tests are most advantageous in situations where the group tested has a range of proficiency too wide to be measured effectively by a conventional linear test with a concentrated difficulty level (Kim & Plake, 1993; Lord, 1971, 1980). Multistage adaptive tests often lead to reduced testing and score reporting time and have been shown to provide equal or higher predictive and concurrent validity of score inferences compared to linear tests (Linn, Rock, & Cleary, 1969; Wainer, 1995; Weiss, 1982). Computerized tests also provide more flexible scheduling for testing.

What are the relative advantages and disadvantages of a multistage test compared to an item-level CAT?

There are several advantages of multistage tests when compared with item-level adaptive tests, and few disadvantages. Because the testlets can be designed and assembled before administration and are scored as a unit, test developers have more control over the quality of the structure and administration of the final test when compared with an item-level CAT. This control eliminates many of the disadvantages associated with item-level CATs cited previously and results in better quality tests including: (1) better assurance of local independence between the testlets and thus the unidimensionality of the test, (2) increased control over item ordering and context effects, (3) control over nonstatistical properties including content specifications, (4) greater control over test security, (5) allowance for some item review, and (6) fewer data management and computer processing demands.

Local independence between the testlets, and thus the unidimensionality of the multistage test composed of testlets, is better assured compared to item-level adaptive tests (Thissen et al., 1989). This is because items that are related to each other by a common stimulus or by similar content are treated as one polytomous item and independence of responses within the testlet is not required. Although local item dependence among items in a testlet is not necessarily eliminated, performance on that set of items in relation to other test items is more accurately measured when a polytomous model is used (Yen, 1993). Testlet adaptive tests have also been shown to provide more accurate (often lower) estimates of the reliability of test scores compared to item-level CAT estimates because local item dependence often exists on item-level CATs. A multistage test that treats locally dependent items as a set/unit and uses a polytomous IRT model provides better (lower) estimates of item discrimination, thus leading to more accurate (lower) estimates of precision and reliability.

The use of item sets may not eliminate context effects but should reduce this possibility. Within stage-based CATs, each item is embedded in a specific item set among other predetermined items. Thus, each item's context is limited, both in a given form and across forms, allowing test developers to carefully scrutinize the item sets and all possible combinations of the item sets across forms before administration, ensuring desired characteristics such as ordering of the items, and excluding negative effects, such as dependencies among the items.

Tests composed of item sets may be especially useful if there are many content areas or complicated cross-classification of items. Because the item sets are constructed before administration, test developers may check in detail that formal content specifications are met as

well as that the informal nonexplicit content characteristics of items are appropriately represented and distributed. For example, it might be too difficult for computer algorithms to check for inappropriate subject matter incidental to content, such as too many items about photosynthesis. Test developers may make these checks easily, however, as well as complete cross-classification of content by item difficulty. Thus, the increased quality control for testlet-based CATs can increase the probability that examinees of similar proficiency receive similar content (Wainer, 1990). Creation of the item set ahead of time also allows for checking the distribution of other nonstatistical properties of the items such as the cognitive level, item format, word count and answer key position; characteristics that were essentially ignored in initial conceptions of CAT.

Item and test exposure are limited and controlled in a multistage test compared to an item-level CAT because the test developers can limit the use of items in different item sets. If testing is limited to a few sessions then one instantiation of a multistage test may be administered per testing session, with parallel multistage tests administered at other sessions, further limiting item exposure. Thus, item exposure in multistage testing is controlled through preconstructing and limiting the number of adaptive test forms, rather than through item exposure control methods and large item pools, as in item-level CATs.

A multistage test with linearly administered stages that are adaptively chosen allows examinees to preview and review items within a stage and to change their answers. Because adaptation only takes place between stages, item review and change within stages does not present difficulty for the administration algorithm or vulnerability to response strategies that allow examinees to maximize their scores (Vispoel et al., 2000).

Finally, the reduced number of adaptation points in a multistage test may lead to quicker scoring and reduced demands for routing, data management, and computer processing when compared with an item-level CAT, all potentially contributing to a more efficient test administration (Wainer, 1990).

While there are many advantages of multistage tests compared to item-level CATs, there are also a few limitations. The primary limitation is that generally more items are needed to get the same measurement precision. Additionally, construction of multistage tests may require more work on the part of item writers and test editors than an item-level adaptive test. The test developers must carefully scrutinize the item sets before administration to ensure desired characteristics and exclude negative effects, including content distribution by item difficulty, dependencies among the items, and distribution of nonstatistical properties of the items such as the cognitive level, item format, word count and answer key position. Finally, it may be difficult to replace items of a testlet independently of the others, as the items are treated as a unit within the testlet (Wainer & Kiely, 1987).

Two-stage tests have the added disadvantage of a higher likelihood of routing errors due to the one adaptation point. This likelihood is especially high for examinees whose scores fall near the routing-test cut scores, and these errors may be exacerbated if examinees are guessing. Using more stages may guard against these errors (Weiss, 1974). Alternatively, a 'recovery routine' may be built in for cases in which it is obvious that the incorrect second-stage test was chosen or if the test-taker's proficiency is between the difficulty levels of two adjacent tests.

In sum, using item sets allows for test developers to have more control over the test process and quality and provides many advantages and relatively few disadvantages when compared with item-level adaptive and linear tests. Multistage tests seemingly combine the advantages of adaptive testing with the advantages of linear tests (Berger, 1994), providing a balanced compromise between these two test forms. This has led to the increased research and use of multistage tests.

Multistage Tests in Practice

Multistage testing has taken several forms and names, including computerized mastery testing (CMT) (Lewis & Sheehan, 1990), Computer-adaptive sequential testing (CAST) (Luecht & Nungester, 1998), multiple form structures (MFS) (Armstrong, Jones, Koppel, & Pashley, 2004), and Bundled Multistage Adaptive Testing (BMAT) (Luecht, 2003). In particular, the CAST framework has been used for the U.S. Medical Licensure Examination and the MFS method has been applied in a computerized version of the Law School Admission Test. The developers of these tests use different terminology and made different choices during development, but they each had to address questions about the number and length of the stages/testlets, and the target statistical and qualitative characteristics of the test, the stages, and the testlets as discussed in the next sections.

Issues in Developing Multistage Tests

Creating a multistage test requires many of the same decisions as creating a linear test or an item-level CAT. How long will the test be? What content will the test cover? What is the desired difficulty for the entire test and how will this be achieved? While it is important to consider these questions at the entire test level, these questions must also be considered at the testlet level. Specifically, how long will each testlet be? What content will each testlet cover? What is the desired difficulty for each testlet? How many testlets will be used at each stage? How will the testlets be scored? How will the testlets be chosen? The following are just some of the steps to be considered throughout the process of creating a multistage test.

First, as with any test development, the purpose of the test must be determined. The population to be tested and the decisions to be made from the test must also be considered. These considerations will help to determine the length of the test, whether an adaptive stopping rule will be used the content coverage of the entire test, and the necessary difficulty distribution of the item pool. From these considerations, a table of specifications can then be developed.

Next, items need to be developed to cover the desired content, difficulty, and information range for the total test. The item pool must support the assembly of multiple multistage tests. Then the items need to be assembled into testlets and stages of the test. This test assembly process will be guided by several questions. The solutions chosen will depend on the purposes of the testing program and the particular test under consideration. These questions, listed in Table 1, as well as guidelines and recommendations for addressing them based on previous research and use, are discussed in further detail below.

Table 1. Questions to be Answered in Building a Multistage Test

How many stages?
How many testlets at each stage?
How long should the testlets be?
What are the target statistical and qualitative specifications within and across testlets and stages?
How are the items/stages scored?
When and how to adapt?
How to assemble the tests?

1. How many stages?
   The possible number of stages ranges from two to the total number of test items. Most research and application has used three or four stages. More stages and more variety of difficulty of testlets within the stages allows for greater adaptation and thus more flexibility. Researchers should keep in mind, however, that adding more stages to the test increases the complexity of the test assembly, without necessarily adding much to the measurement precision of the final test forms (Luecht & Nungester, 1998; Luecht, Nungester, & Hadidi, 1996).
2. How many testlets should be developed for each stage?
   Most research and application has used one testlet in the first stage (routing test) and then increasing numbers of possible testlets for administration across subsequent stages, usually no more than eight and averaging around five. As for the number of stages, adding more testlets and greater variety of difficulty of the testlets allows for greater adaptation and thus more flexibility with the test. Researchers should keep in mind, however, that adding more testlets also increases the complexity of test assembly, without necessarily adding much to the measurement precision of the final test forms (Luecht & Nungester, 1998; Luecht et al., 1996). In general, research indicates that a maximum of four testlets is desirable at any one stage and that three levels may be adequate (Armstrong, et al., 2004).
3. How long should the testlets be?
   Research studies and operational testlet-based tests have used between 1 and 90 items per testlet, with an average of about 5 items per testlet. The length of the testlets may vary across the stages. Some tests have longer first stage (routing) tests and shorter testlets in subsequent stages. Kim and Plake (1993) found that increasing the length of the first-stage (routing) testlet was most important in reducing the size of the proficiency estimate errors. In general, shorter testlets allow for greater efficiency as there is more possibility for adaptation, given a particular number of total test items.
4. What are the target statistical and qualitative specifications?
   A test developer can impose constraints on item selection either at the stage/testlet level or at the complete test level, depending on whether constraints are desired for each separate stage of the test or on the entire multistage test. In either case, one must jointly consider the content and statistical constraints and find the balance between these target characteristics.
   The choice of the statistical targets is one of the most important decisions in designing the testlet-based test. The test developer must determine the desired average item difficulty and range of difficulty that each testlet will cover. If using an IRT model, the target test(let) information and target test(let) characteristic curves must be determined. Generally, the goal is to select the items that maximize the information in the test, subject to the constraint that the test information function continues to reflect the desired shape.
   Kim and Plake (1993) found that the statistical characteristics of the first-stage (routing) testlet had a major influence on the complete test's measurement precision. A routing test with a rectangular-shaped distribution (wide range) of item difficulty parameters provided better measurement at the ends of the ability distribution while a peaked item difficulty routing test (concentrated at a particular level) was better in the middle of the ability distribution, depending on the number of second-stage testlets used.
   The test developer may consider content balancing, context effects, dependencies among the items, cognitive level, item format, word count, answer key position, diversity usage, and any other characteristics of interest or concern in developing the testlets. For example, testlets may be created that are content-balanced or content-specific, but of different difficulty levels.
5. How are the items/stages to be scored?
   Number-correct or IRT trait estimates may be used to score the items for adaptation. If using IRT, one must determine the appropriate model. Research and applications of multistage tests have often used the three-parameter logistic dichotomous model (Birnbaum, 1968) or the nominal (Bock, 1972) or graded response (Samejima, 1969) polytomous models. A particular benefit of multistage tests is that the preconstructed testlets may be best treated as polytomous items and may be scored using a polytomous IRT model. Such a model allows for dependencies among the questions within the testlets, but requires independence between the testlets.
   More recently, Wainer, Bradlow, and Du (2000) developed Testlet Response Theory, an extension of the three-parameter logistic model. This model is advantageous for cases when there is concern about losing information from the entire response pattern when treating each testlet as a polytomous item, or if testlet items are to be adaptively chosen during administration.
6. When and how to adapt?
   Efficient adaptation requires determining the most effective item set to administer at each stage. The closest thing to an item-level CAT, and consequently most efficient, is a multistage test in which adaptation takes place within as well as between the stages or testlets. Increasing adaptation points, however, lends itself to the same issues that face item-level CATs. Furthermore, it has been shown that the gain in efficiency from adapting within testlets is modest (Wainer

et al., 1990, 1992). Thus, current usage of multistage tests generally involves adaptation only between the stages. Identifying which testlet in the next stage is best matched to the examinees' ability estimates is often achieved through maximizing information or minimizing standard error based on performance on the previous stage(s).

7. How to assemble the tests?
Different forms of the multistage test must be assembled. This may be achieved through an automatic test assembly (ATA) process that uses mixed-integer programming to minimize linear objective functions and allows for a large variety of constraints. Mathematical procedures and commercial computer software exist for these ATA procedures. See Adema (1990), Luecht & Nungester (1998), and van der Linden & Adema (1998) for more information regarding ATA.

Summary

Research with and use of non-item-level adaptive tests, including two-stage or testlet-based, has revealed the potential advantages of these models. These types of adaptive tests allow for more efficient measurement of more examinees compared to linear tests, while protecting against some of the problems encountered with item-level adaptive tests. The use of stages and testlets allows the knowledge and skills of the test developers into the process of test development, rather than simple reliance on the statistical characteristics of items (e.g., information) to construct the test (Wainer, 1990; Wainer & Kiely, 1987). As Wainer and Kiely (1987) summarized, ". . . [multistage] testlets are a scheme which can maintain the CAT advantages while still using the wisdom of experts."

Acknowledgements

The author would like to thank Deborah Harris, Brandi Weiss, and three anonymous reviewers for their feedback on earlier versions.

References

Adema, J. J. (1990). The construction of customized two-stage tests. *Journal of Educational Measurement*, *27*, 241–253.

Armstrong, R. D., Jones, D. H., Koppel, N. B., & Pashley, P. J. (2004). Computerized adaptive testing with multiple-form structures. *Applied Psychological Measurement*, *28*, 147–164.

Berger, M. P. F. (1994). A general approach to algorithmic design of fixed-form tests, adaptive tests, and testlets. *Applied Psychological Measurement*, *18*, 141–153.

Betz, N. E., & Weiss, D. J. (1973). *An empirical study of computer-administered two-stage ability testing*. Psychometric Methods Program, Research Rep. No. 73–4, Department of Psychology, University of Minnesota, Minneapolis.

Birnbaum, A. (1968). Statistical theory for logistic mental test models with a prior distribution of ability. *Journal of Mathematical Psychology*, *6*, 258–276.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*, 29–51.

Bock, R. D., & Zimowski, M.F. (1998). *Feasibility studies of two-stage testing in large-scale educational assessment: Implications for NAEP*. Commissioned by the NAEP Validity Studies Panel, American Institute for Research in the Behavioral Sciences. Washington, D.C.: NCES.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer Nijhoff Publishing.

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

Kingsbury, G. G. (1996). *Item review and adaptive testing.* Paper presented at the Annual Meeting of the National Conference on Measurement in Education, New York.

Lewis, C., & Sheehan, K. (1990). Using Bayesian decision theory to design a computerized mastery test. *Applied Psychological Measurement*, *14*, 367–386.

Linn, R. L., Rock, D. A., & Cleary, T. (1969). The development and evaluations of several programmed testing methods. *Educational and Psychological Measurement*, *29*, 129–146.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.

Loyd, B. H. (1984). *Efficiency and precision in two-stage adaptive testing*. West Palm Beach, FL: Eastern ERA.

Luecht, R. M. (2003). *Exposure control using adaptive multistage item bundles.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago.

Luecht, R. M., & Nungester, R. J. (1998). Some practical examples of computer-adaptive sequential testing. *Journal of Educational Measurement*, *35*, 229–249.

Luecht, R. M., Nungester, R. J., & Hadidi, A. (1996). *Heuristic-based CAT: Balancing item information, content and exposure.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, New York.

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores (Psychometric Monograph No. 17)*. Iowa City, IA: Psychometric Society.

Schnipke, D. L., & Reese, L. M. (1997). *Comparison of testlet-based test designs for computerized adaptive testing*. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago.

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement*, *28*, 237–247.

Thissen, D., Steinberg, L., & Mooney, J. A. (1989). Trace lines for testlets: A use of multiple-categorical-response models. *Journal of Educational Measurement*, *26*, 247–260.

van der Linden, W. J., & Adema, J. J. (1998). Simultaneous assembly of multiple test forms. *Journal of Educational Measurement*, *35*, 185–198.

van der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of modern item response theory*. New York: Springer.

Vispoel, W. P. (1998). Reviewing and changing answers on computer-adaptive and self-adaptive vocabulary tests. *Journal of Educational Measurement*, *35*, 328–345.

Vispoel, W., Hendrickson, A. B., & Bleiler, T. (2000). Limiting answer review and change on computerized adaptive vocabulary tests: Psychometric and attitudinal results. *Journal of Educational Measurement*, *37*, 21–38.

Wainer, H. (1990). *Computerized adaptive testing: A primer*. Hillsdale, NJ: Lawrence Erlbaum.

Wainer, H. (1993). Some practical considerations when converting a linearly administered test to an adaptive format. *Educational Measurement: Issues and Practice*, *12*(1), 15–20.

Wainer, H. (1995). Precision and differential item functioning on a testlet-based test: The 1991 Law School Admissions Test as an example. *Applied Measurement in Education*, *8*, 157–186.

Wainer, H., Bradlow, E. T., & Du, Z. (2000). Testlet response theory: An analog for the 3PL model useful in testlet-based adaptive testing. In W. J. van der Linden and C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 245–269). Dordrecht, the Netherlands: Kluwer Academic Publishers.

Wainer, H., Kaplan, B., & Lewis, C. (1992). A comparison of the performance of simulated hierarchical and linear testlets. *Journal of Educational Measurement*, *29*, 243–251.

Wainer, H., Lewis, C., Kaplan, B., & Braswell, J. (1990). *An adaptive algebra test: A testlet-based, hierarchically-structured test with validity-based scoring*. ETS Technical Report 90–92. Princeton, NJ: ETS.

Wainer, H., & Lukhele, R. (1997). How reliable are TOEFL scores? *Educational and Psychological Measurement*, *57*, 741–758.

Weiss, D. J. (1974). *Strategies of adaptive ability measurement*. Psychometric Methods Program, Research Report 74–5, Department of Psychology, University of Minnesota, Minneapolis.

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement*, *6*, 473–492.

Wise, S. L. (1996). *A critical analysis of the arguments for and against item review in computerized adaptive testing*. Paper presented at the Annual Meeting of the National Conference on Measurement in Education, New York.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, *30*, 187–214.

ANNOTATED REFERENCES

Kim, H., & Plake, B. S. (1993). *Monte Carlo simulation comparison of two-stage testing and computerized adaptive testing*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Atlanta. A study comparing eighteen simulated two-stage and item-level adaptive tests. Shows how characteristics of the first-stage test influence measurement precision of ability. Reviews limitations of item-level CATs and advantages of two-stage tests.

Lord, F. M. (1971). A theoretical study of two-stage testing. *Psychometrika, 36*, 227–242. Discusses considerations in designing a two-stage test and use of the information function for measuring the effectiveness of the test procedure. Provides theoretical results for 5 testing procedures—both two-stage and item-level adaptive tests. Discusses the affects of the test characteristics including possibility of guessing, length of first-stage tests, number of second-stage tests, and difficulty of stages.

Lord, F. M. (1974). *Practical methods for redesigning a homogeneous test, also for designing a multilevel test*. Educational Testing Service RB-74–30. Presents practical methods for designing a multi-level test. Reviews important characteristics of a multistage test including the number and relationship of stages and how to route examinees to levels.

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24*, 185–201. Reviews the development of computerized adaptive testing and the limitations of such tests, including context effects, lack of robustness, and item difficulty ordering. Then develops a model for testlets, both hierarchical and linear, and discusses the relative advantages of testlet-based adaptive tests.

Self-Test

1. What is a multistage test?
2. What is a testlet?
3. What are the relative advantages and disadvantages of a multistage test compared to a paper-and-pencil test?
4. What are the relative advantages and disadvantages of a multistage test compared to an item-level CAT?
5. What are the considerations in creating a multistage CAT?

Answers to Self-Test

1. A multistage test consists of sets of items that are scored as a unit and that are adaptively administered to examinees. Thus, scores from the first-stage tests are used to determine which second-stage test each examinee should complete to obtain the most precise measurement. This choice is made from several alternative second-stage, or measurement, tests containing items concentrated at a particular level of difficulty. This process continues in a multistage test such that examinees are routed to tests with more narrowly focused difficulty levels at higher stages based on their performance on previous stages.
2. Wainer and Kiely (1987) described a testlet as a group of items that relate to a single content area, are constructed and analyzed as a unit, and are statistically independent of all other testlets and items. Testlets are created by test specialists before the adaptive test is administered and are scored as polytomous items. Testlets are often used for questions that are related to a common stimulus, such as reading comprehension passages. Another commonly used type of testlet is one that is content-balanced or content-specific. A general definition of testlets allows for the items to be stimulus-dependent, content-balanced, content-specific, or any other set of items, each of which are built to a particular range of ability.
3. Advantages of multistage test compared to a paper-and-pencil test:
    a. More efficient and precise measurement across the proficiency scale.
    b. Often leads to reduced testing and score reporting time.
    c. Has been shown to provide equal or higher predictive and concurrent validity of score inferences.
    d. Computerized versions provide more flexible scheduling for testing.
4. Advantages of multistage test compared to an item-level CAT:

    Advantages:

    Allow for more control over the administration and structure of the final tests including:

    i. Better assurance of local independence between the testlets and thus the unidimensionality of the test
        a) Provide more accurate (often lower) estimates of the reliability of test scores
    ii. Increased control over item ordering and thus reduced possibility of context effects and other item dependencies
    iii. Control over non-statistical properties including content specifications, cognitive level, item format, word count and answer key position
    iv. Item and test exposure are limited and controlled
    v. Allows examinees to preview and review items within a stage and to change their answers
    vi. Fewer data management and computer processing demands

    Disadvantages:

    a. Generally more items are needed to get the same measurement precision.
    b. May require more work on the part of item writers and test editors:
        i. Must ensure desired characteristics and exclude negative effects
        ii. Must check for inappropriate subject matter incidental to content and complete a cross-classification of content by item difficulty
        iii. Must check for dependencies among the items

iv. Must check distribution of non-statistical properties of the items such as the cognitive level, item format, word count and answer key position

c. May be difficult to replace items of a testlet independently of the others, as the items are treated as a unit within the testlet.

d. Two-stage tests have the added disadvantage of higher likelihood of routing errors.

5. How long will the test be? What content will the test cover? What is the desired difficulty for the entire test and how will this be achieved? How long will each testlet be, how many testlets will be used, what content will each testlet cover, what is the desired difficulty for each testlet, and how will the testlets be chosen? How will scoring and adaptation take place within the test? Will adaptation occur within the testlets or only between? How will the tests be assembled?