Evaluating the Content Validity of Multistage-Adaptive Tests

Katrina Crotts, Stephen G. Sireci, and April Zenisky

University of Massachusetts Amherst

Conflict of interest statement: This research was supported in part by a contract with the Massachusetts Department of Elementary and Secondary Education, office of Adult and Community Learning Services. The opinions expressed in this paper are those of the authors and do not represent official positions of the Massachusetts Department of Elementary and Secondary Education.

Correspondence concerning this manuscript should be sent to Stephen G. Sireci, Center for Educational Assessment, University of Massachusetts, Amherst, MA 01003. Email correspondence may be sent to Sireci@acad.umass.edu.

Abstract

Validity evidence based on test content is important for educational tests to demonstrate the degree to which they fulfill their purposes. Most content validity studies involve subject matter experts (SMEs) who rate items that comprise a test form. In computerized-adaptive testing, examinees take different sets of items and test “forms” do not exist, which makes it difficult to evaluate the content validity of different tests taken by different examinees. In this study, we evaluated content validity of a multistage-adaptive test (MST) using SMEs’ content validity ratings of all items in the MST bank. Analyses of these ratings across the most common “paths” taken by examinees were conducted. The results indicated the content validity ratings across the different tests taken by examinees were roughly equivalent. The method used illustrates how content validity can be evaluated in an MST context.

Evaluating the Content Validity of Multistage-Adaptive Tests

Validity evidence based on test content is a fundamental requirement for defending the utility of an educational test for its intended purposes (American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME), 1999). Such evidence has traditionally been called content validity evidence, although some validity theorists question the use of this term because evaluations of test content focus on the test, rather than on interpretations of scores (e.g., Messick, 1989). However, since content-related validity evidence supports and clarifies test interpretations, it is clear that any evidence bearing on the appropriateness of the test content has important implications for the validity of score interpretations (Sireci, 1998).

Evidence based on test content is typically evaluated using subject matter experts (SMEs) either in the form of alignment studies, or more traditional content validity studies (Martone & Sireci, 2009). In these studies, SMEs rate items comprising a test form with respect to how well they measure specific aspects of the domain of interest. In a more traditional content validity study, the aspects that are rated tend to be item congruence, where SMEs match test items to specified content or skill areas; or item relevance, where SMEs rate the relevance of items to specific objectives, benchmarks, or other elements in the test specifications. Newer evaluations of test content based on alignment methodology also involve rating aspects such as cognitive challenge and examining the agreement between items and the achievement levels or other score interpretations.

Regardless of the approach taken, the similarity among the more traditional and alignment-based evaluations of test content is probably best described as a low complexity alignment study (Bhola, Impara, & Buckendahl, 2003). Although alignment and more traditional content validity studies use different terms (e.g., item-objective congruence, categorical concurrence, content centrality, etc.), they often address the same critical validity question—how well do the items administered to examinees adequately represent the knowledge and skill domain purportedly measured?

In traditional paper-and-pencil assessments, both content validity and alignment studies are conducted by evaluating test items on fixed test forms. However, in a computerized-adaptive context, examinees receive different items tailored to their proficiency level and so no specific test "form" exists, making it difficult to properly evaluate the quality and appropriateness of the test content administered to an examinee. Although most adaptive tests contain item selection constraints to ensure adherence to content specifications, individual examinees still experience relatively unique sets of items making it difficult to evaluate the consistency of content validity across all possible test "forms." As Luecht (2005) pointed out, the lack of pre-constructed forms in an adaptive context means less ability to conduct quality control in general, suggesting many computer-based testing programs may fail to properly evaluate the content validity of the different sets of items taken by examinees.

In this study, we evaluated the content validity of the Massachusetts Adult Proficiency Test (MAPT) for Reading taken by adult education students. The MAPT is a multistage computerized-adaptive test administered over the Internet to adult education students in Massachusetts, as part of the Federal accountability program in adult education programs in the United States (Sireci et al., 2008). Because the MAPT is adaptive, students see different sets of items, so the idea of a test "form," which is a necessity in paper-based testing, does not apply. Instead, students take an initial set of items, and then are routed to subsequent item sets based on how well they performed on the items taken thus far.

Multistage-Adaptive Testing

In a multistage-adaptive test (MST), the test adaptation occurs based on cumulative performance on a set of items, rather than on performance at each individual item as is done in a traditional (i.e., item-level) computerized-adaptive test (CAT) design. Similar to a CAT, if the examinee is doing well, a more difficult set of items is administered, and if the examinee is not doing so well, an easier set of items is administered (Wainer, 1993). MSTs offer an advantage over CATs in that they offer control over “test assembly and test form quality control, exposure of test materials, facilitating data management, and reducing requirements for test delivery software to handle complex scoring and item selection algorithms” (Luecht, Brumfield, & Breithaupt, 2006, p. 200).

The design of an MST is defined by the number of stages within the test. These stages are represented by sets of items called “modules” or “testlets.” For example, if the MST design were 1-3-3, there would be a total of 3 stages. In the first stage, there is one module, typically referred to as the routing module or locator test (Hendrickson, 2007), and this first-stage module is composed of items that are of moderate difficulty. All examinees begin the test with this module, which is used to determine which of the three modules will be administered to the examinee in stage two, where the modules vary from one another on the basis of difficulty (typically easy, medium, and hard). The final stage also has three modules, similarly differentiated on difficulty, and the third-stage module to be presented to each examinee is determined based on the examinee’s performance in stage two. The design of the MAPT for Reading is more complicated than this “1-3-3” example, and is described in the method section.

Research on Content Validity in an Adaptive Context

Limited research has focused on evaluating content validity in an adaptive context. Luecht, De Champlain, and Nungester (1998) examined the effects of ignoring content balancing in the adaptive testing algorithm. Two simulation studies were conducted to examine the problems related to both content and statistics in a CAT context. Each simulation study involved balanced and unbalanced tests. The unbalanced tests were administered based on item information, whereas the balanced tests were administered based on specific content restrictions while also maximizing information. Results across both simulations revealed no impact in score distributions between the two tests. However, on the unbalanced assessments, an examinee's proficiency level impacted the items that were seen, with lower performing examinees seeing items in mostly one content area and higher performing examinees seeing items mostly of another content area. If the intended purpose of the test is to show adequate knowledge of both content areas, the test score interpretations would not be appropriate for this assessment.

Kaira and Sireci (2010) evaluated content validity on a multistage-adaptive test, the Massachusetts Adult Proficiency Test in Math, focusing specifically on the degree to which the items were measuring their intended content and cognitive areas. SMEs reviewed items and test specifications and made two ratings for each item. The first rating involved selecting the content area measured by the item; the second rating involved selecting the cognitive level measured by the item. Although they found consistency across the paths and panels in the multistage test with some exceptions, they suggested that future research should "investigate more comprehensive measures of item quality, such as the degree to which the item measures its intended benchmark along an ordinal rating scale" (p. 23), rather than simply judging which content areas and cognitive levels each item measured.

Controlling for Content in Adaptive Testing

Item selection in an adaptive context is done through the use of a sophisticated algorithm. Depending on the assessment, the algorithm attempts to balance content coverage, item exposure, and measurement precision (Huff & Sireci, 2005; Luecht, 2005). If the item selection algorithm includes content constraints, content validity is thought to be "built-in" to the item selection algorithm because all items in the item bank are coded with respect to content and cognitive attributes. Thus, some test developers may argue that evaluating content validity in adaptive testing merely means ensuring the item selection algorithm properly meets the content specifications for the test. Additionally, in many cases there is an assumption that because the item selection algorithm is in place, the different test forms for different examinees will be comparable in terms of content and cognitive representation (Kaira & Sireci, 2010).

The argument that content validity is built into a CAT or MST algorithm is insufficient because it assumes the content area and cognitive level codes assigned to each item are infallible. Although an item selection algorithm might ensure adequate *content representation*, it is important to note that content validity is not *binary* (Huff & Sireci, 2005). That is, items measuring a given educational objective may have equivalent content codes, but the degrees to which the items appropriately measure the objective are not likely to be equal. On an adaptive test, examinees can take tests that meet content specifications, but have items of varying content quality, such as the degree to which an item adequately measures its intended knowledge or skill.

Study Purpose

The purpose of this study was to evaluate content validity of a multistage-adaptive test using a more comprehensive measure of item quality than the binary content codes used in previous research. Using SMEs' ratings of the degree to which items measured their intended benchmarks, we investigated the following questions,

1. Are there differences with respect to the content validity of the MAPT items taken by different examinees on different test "forms"?
2. Are there differences with respect to content validity across the different "types" of paths examinees take?

By "types of paths" we were interested in comparing the content validity of tests taken by examinees who were not routed to different levels while taking the test to those who were routed to different proficiency levels while taking the test. Examining these questions provides insight into some of the issues and methods surrounding the evaluation of content validity in an adaptive context.

Method

Data

The MAPT for Reading is a multistage-adaptive test designed for adult education students in Massachusetts. It is aligned with the reading standards and benchmarks in the *Massachusetts Adult Basic Education Curriculum Framework for the English Language Arts* (Adult and Community Learning Services, 2005). The purpose of the MAPT for Reading is to measure the knowledge and skills of Massachusetts Adult Basic Education (ABE) learners in reading and evaluate whether ABE learners are meeting their educational goals (Sireci et al., 2008).

The version of the MAPT for Reading analyzed in this study comprised 320 items arranged into eight sets of vertically equated 40-item tests. The MAPT for Reading involved a 6-stage MST design organized by modules (sets of items) and panels (an entire set of modules that represents all possible sets of items that could be taken by an examinee). An illustration of the MAPT design is presented in Figure 1. At the time of this study, the MAPT for Reading had four modules within each stage, and was made up of two parallel panels[1]. Two panels, identical in design, were developed because students are required to take the test twice. Thus, these panels represented "parallel forms" of the entire MAPT for Reading.

[Insert Figure 1 about here]

As seen in Figure 1, there are four possible entry points (Levels 1 to 4), which represent different "educational functioning levels" (similar to grade levels in K-12 education). These different entry points are needed to account for the different levels of proficiency that exist for the adults who take these assessments. The first time a student takes the MAPT, their classroom teacher decides their entry point. Thereafter, their most recent MAPT score is used to determine their entry point (Sireci et al., 2008). Although we refer to these levels as 1 through 4, the National Reporting System (NRS) for Adult Education defines them as beginning basic, low intermediate, high intermediate, and low adult secondary.

The data from this study come from two parallel MAPT for Reading panels (Panel A and Panel B), each consisting of four, 40-item tests (one at each level) that were in place from July 1, 2009 through June 30, 2010. Thus, each panel comprised 160 items. There were 6 stages, with Stage 1 consisting of 15 items, and Stages 2-6 each having 5 items, with no items being duplicated across modules or panels. To move across stages, an item response theory (IRT)-based estimate of Reading proficiency is calculated, and this ability estimate is used to determine the next module administered (Sireci et al., 2008). This movement through the different stages is called a path, which essentially represents a test "form" for each examinee. The arrows in Figure 1 indicate some of the possible paths that an examinee can take; however, although not shown in

[1] Currently, the MAPT has five modules per stage and three panels.

Figure 1, it is possible to "travel" to non-adjacent modules. This means an examinee could start at Level 1 and then be routed to Level 3 in Stage 2. Given the fact that there are a total of 24 modules (6 Stages x 4 Levels) in each panel, the potential number of test "forms" for the MAPT is very large.

The MAPT for Reading was administered to approximately 8,600 adults in Massachusetts during the 2010 fiscal year (July 2009 through June 2010). Analyzing all possible paths examinees took across the 6 stages and 4 levels would be too impractical so we decided to select a subgroup of paths that would best represent the entire set of observed paths. The "straight paths," which referred to the situation where a student started at a particular test level and remained in that level for the 5 subsequent stages, were among the most popular and so they were selected for analysis. The straight paths are of particular interest since the test specifications are built to represent the content and cognitive targets specified at each level. Next, we selected the next most popular paths taken by examinees until we obtained at least half of the actual paths taken during the fiscal year. A total of 14 paths were examined in *each* panel. Although there were many other paths traveled by examinees, they involved much smaller sample sizes. The 14 paths selected were representative of the different types of paths taken by examinees and, as described next, included straight paths, and paths where examinees were routed to more difficult or less difficult modules.

Table 1 presents the number of students taking each of the 14 specific paths in each panel. These frequencies ranged from 130 to 660. The straight paths (e.g., 1-1-1-1-1-1) were the most common paths traveled, accounting for about 30% of all test administrations. A variety of "mixed paths" were also observed, including paths where students increased one level (e.g., 3-4-4-4-4-4), decreased one level (e.g., 2-1-1-1-1-1), or experienced both increases and decreases in level across the stages (e.g., 2-3-2-2-3-2). Only one of the 14 paths involved students being routed more than one level (i.e., examinees started at level 2, were routed to level 4 at the second stage and remained there for all subsequent stages). The 14 paths used in this study appear in both Panels A and B, meaning we examined 28 different paths, which are analogous to test forms taken by examinees. In Table 2, we provide the classification types for each of the 14 paths.

Note that the unit of analysis in this study refers to the number of items, not the number of examinees. Each MAPT test administration consisted of 40 items. Thus, there were 320 unique items across the two panels (40 Items x 4 Levels x 2 Panels). As part of an earlier content validity study, six subject matter experts (SMEs) in adult education with at least three years of teaching experience were convened to review the items and the benchmark to which the item was written. Specifically, each SME was asked to “rate how well the item measures its intended benchmark” using a scale of 1 to 6 where 1= “not at all” and 6= “very well.” Intermediate values on the scale were not defined. The panelists were oriented to the purpose of the study and were trained on the rating task using practice items. More complete details of the study can be found in Zenisky et al. (2012).

For the purposes of this study, the ratings were averaged across the SMEs to provide a content validity index for each item. These mean ratings served as our index of content quality and served as the dependent variable in our analyses[2]. For the most part, only items meeting a minimum standard of adequacy were selected for operational use. Nevertheless, there was variability across the operational items with respect to how well they measured the intended benchmarks.

---

[2] Although the original ratings come from an ordinal scale, the means and medians were very similar across items and so we used the mean rating to provide a dependent variable that had somewhat greater variability.

Some older items were not included in the content validity study because they were selected for retirement at the end of the testing cycle due to changes in the test specifications and were not rated by the SMEs. These items (39 of 320, about 12%) were excluded from the analyses and were not linked to any particular content or cognitive area. Although each test administration is comprised 40 items, due to the elimination of items not rated by SMEs, the actual number of items per path analyzed here ranged from 33 to 39.

Data Analysis

Descriptive statistics and analysis of variance (ANOVA) were used in making comparisons to address the research questions. First, we were interested in examining whether there were differences across the "straight paths" within and across panels. This analysis involved a two-way ANOVA using path (4 levels) and panel (2 levels) as independent variables, and the content validity index as the dependent variable. Next, we wanted to compare all of the 14 most popular paths within and across panels. This analysis also involved a two-way ANOVA with path having 14 levels and panel having two levels. However, it should be noted that this analysis involved items that overlapped since the same modules could be involved in different paths. Thus, the power of the ANOVA results is weakened by this partial dependence that was not accounted for in the model. Third, we were interested in examining the differences in item quality with respect to content quality across the four most popular path "types" including straight, increasing, decreasing, and increasing-decreasing. For this analysis, we conducted a one-way ANOVA with planned comparisons. Although these planned comparisons may seem like a subset of the 14-path ANOVA, there were more items within each path *type* as opposed to individual paths, therefore more degrees of freedom and more power to evaluate content differences across paths.

Results

Table 3 presents descriptive statistics for the mean content validity indices for the items in each path and panel. Given the 1-6 scale on which these indices are based, the average ratings are consistently high for each 40-item test, ranging from 4.5 (path 1-1-1-1-1-1 in Panel A) to 5.4 (path 3-3-3-3-3-3 in Panel A). Additionally, we can see that across the 14 different paths and the two panels, the mean content validity index was quite similar with a mean rating of 5.01 for Panel A, and 5.03 for Panel B.

Based on the results of the descriptive statistics and the similarities in ratings across the different paths and panels, the results of the two-way ANOVA were not surprising (see Table 4). The results for the straight path-by-panel two-way ANOVA were not statistically significant for the main effect of path ($F_{(3,277)}$=2.1, p=0.10), panel ($F_{(1,277)}$=0.01, p=0.91), or path-by-panel interaction ($F_{(3,277)}$=0.9, p=0.45). The results of the second ANOVA (for the 14 paths by two panels) were also not statistically significant for the main effect of path ($F_{(13,986)}$=1.2, p=0.24), panel ($F_{(1,986)}$=0.1, p=0.81), or path-by-panel interaction ($F_{(13,986)}$=0.7, p=0.77). These findings indicated that the content validity, as measured by mean SME ratings, was consistent across the 14 different paths and two panels.

Descriptive statistics for the tests regrouped as path "types" are presented in Table 5. Note the large increase in the number of items (sample size) per group. The omnibus F for the ANOVA associated with the grouping of paths into types again did not yield statistical significance ($F_{(3,1006)}$=2.2, p=0.09) (see Table 4). Nevertheless, since we were interested in the differences across the different types of paths, we interpreted the significance of the 6 planned comparisons. Two of the planned comparisons were statistically significant at $p < 0.05$. The increasing-decreasing paths had statistically significantly *higher* content validity ratings than both the straight paths (p=0.04) and the decreasing paths (p=0.03). This result is interesting in that examinees who crossed levels more than once ended up with tests that had slightly higher item-benchmark congruence, than the other paths. However, the effect sizes associated with these differences were small (Cohen's d=0.17 and 0.20, respectively).

Discussion

In this study, we used the ratings of SMEs regarding how well items measured their intended benchmark as an index of the content validity of each item within a multistage-adaptive test. We investigated whether the actual tests taken by examinees differed with respect to content validity. Given that test developers do not have total control over the different sets of items taken by examinees in an adaptive context, we were interested to see if this lack of control would affect the overall content validity of the exams. Because multistage tests offer a compromise between item-level computerized-adaptive tests and linear tests, they have become increasingly widespread in both use and in research (Hendrickson, 2007), therefore making it essential to understand how to investigate content validity in an adaptive context.

The results indicated the content validity of the different exams taken by examinees (28 different exams total) were roughly equivalent. The results are good news for the MAPT for Reading, in that we did not discover any inconsistencies in content validity across the subsets of items taken by a majority of examinees. More important, however, is that our analyses illustratre the types of investigations that could be done to evaluate test content within an adaptive context.

Because content validity is assumed to be built into an adaptive test by the item selection algorithm, it has not been widely studied in an adaptive context. Although some studies have examined the important aspects of content balancing (e.g., Luecht et al, 1998) with respect to content validity, other aspects, such as the degree to which an item adequately measures its intended benchmark, should not be ignored. Given that items are likely to differ with respect to degree to which they measure their intended objectives, content validity is not an attribute that can be dichotomously coded into an item selection algorithm. Although content and cognitive codes for item attributes are important aspects of an adaptive algorithm, content validity indices should also be considered in adaptive testing to better guarantee examinees take tests of comparable quality, regardless of which paths they take. That is, in addition to the algorithm keeping track of item difficulty and test specifications, it could ensure examinees see items with a minimum level of content quality.

Thissen and Mislevy (1990) noted that content balancing procedures in an adaptive test may be enforced in a manner similar to exposure control, where the algorithm could that only items of high quality are selected (Thissen & Mislevy, 1990). It is important to note, however, that additional restraints within the algorithm could impact and reduce efficiency and precision (Reese, Schnipke, & Luebke, 1999).

Limitations

Although our analyses have heuristic value, this study had several limitations. First, all the items studied had already been screened for content quality and so there was little room for content irrelevance to appear. Other testing programs with less rigorous screening of item-benchmark congruence may have very different results. Other limitations are the relatively shallow depth of the item bank and the number of modules in the multistage test design. Even though the testing context was adaptive, and the items in each module were unique, we only studied 28 different paths (14 in each panel). Thus, many of the same modules appeared in the different paths and there was overlap with respect to the items that showed up in each path. It would be interesting to evaluate content quality in an item-level adaptive context, or in a multistage test design that featured more modules. In contexts where there is less overlap among the sets of items taken by different examinees (e.g., high-stakes contexts where item security is more critical), there will be greater differences in test content across examinees, and a greater need to study differences in content quality.

Recommendations for Evaluating Content in an Adaptive Context

This study underscores the importance of screening items based on an independent content validity study. Although the test analyzed was an adult education test[3], the methods and issues used can be applied to all situations in which computer-adaptive or multistage-adaptive tests are used. Based on the results and procedures for this study, some recommendations are proposed for test developers evaluating content validity in an adaptive context.

First, before items are placed into the item bank for operational use, they should be pre-screened for content quality. This means carefully running a content validity study with SMEs, and choosing only the items that are rated highly for measuring the intended construct. It is recommended that an ordinal scale (or interval, if possible) be used as a measure for content validity indices because content is not binary and has different degrees of measuring the intended construct. Pre-screening the items would remove any potential for poorly rated items ending up on a test “form.” Pre-screening should also thoroughly check the adaptive testing algorithms to ensure that these algorithms are robust under problem scenarios (Luecht, 2005).

Once items are operational, both quality assurance and control procedures should be in place to ensure that only items with high quality are administered. MSTs offer an advantage over the item-level CAT in that it is easier to preconstruct and prepackage test forms prior to and during test administration to conduct appropriate quality control checks (Luecht, 2005). With an

[3] It should be noted that the MAPT is based on adult education curriculum frameworks that are similar to those found in K12 statewide achievement testing.

MST, modules could be pre-assembled to ensure that each individual module with the test form is of high quality (Reese et al., 1999), thus ensuring that the full test form will also be of high quality.

We understand it may be very difficult to examine all possible paths on an adaptive test, especially item-level adaptive tests. To address this issue, it is recommended that path types be clustered together by path "type," rather than examining all possible paths. Additionally, paths could be stratified by proficiency to look similar to the straight paths used on the MAPT. These paths could then be randomly sampled and examined more closely. If there are significant differences found among path types or among the proficiency paths, then those individual paths within that path type should be further examined.

Lastly, as a method of quality control when the assessment is administered, it is recommended that content quality controls be added to the content constraints within the item selection algorithm. Because content quality should be of sufficient depth to support valid interpretations of test scores (American Council on Education, 1995), adding content quality constraints would aid in valid test score interpretations. Unfortunately, too many constraints can compromise the precision and efficiency of an adaptive assessment (Reese et al., 1999). Although this could mean diminishing returns regarding the use of an adaptive test, it is important to note that adaptive assessments still offer advantages over a traditional linear assessment. Specifically, adaptive assessments offer more appropriate item administration, precise test scores, a flexible test environment, immediate score reporting, and in some cases fewer test questions (American Council on Education, 1995).

Suggestions for Future Research

With the development of more computer-adaptive tests, both item-level and multistage, there is a greater need to study differences in content quality among test “forms.” This study built off of a previous study (Kaira & Sireci, 2010) by examining a more comprehensive measure of content validity. This measure was not dichotomous, but instead measured the degree to which an item measured its intended standard along an ordinal rating scale. However, future research is still necessary to understand how to evaluate content validity in an adaptive context, including item-level adaptive CATs. Additionally, future research should evaluate content quality in an MST design that featured more modules and panels. Specifically, it would be beneficial to examine contexts with less overlap among the sets of items taken by different examinees (i.e., higher stakes tests). Until further studies are done to evaluate the content validity of other adaptive tests, both item-level and multistage, the degree to which content validity varies across tests taken by examinees in these context will remain unknown.

Test construction guidelines and professional standards for testing (e.g., AERA et al., 1999) recommend evaluating items for content quality, and then using those data to select the best (i.e., most aligned or content-valid) items for operational use. We recommend that in an adaptive context, these data also be used in the item selection algorithm to set minimum standards for content quality. Treating content characteristics of items as continuous, or at least ordinal data, rather than nominal data, is likely to promote content validity in adaptive testing situations.

References

Adult and Community Learning Services (2005). *Massachusetts adult basic education curriculum framework for the English language arts*. Malden: MA: Massachusetts Department of Education. Available at http://www.doe.mass.edu/acls/frameworks.

American Council on Education. (1995). *Guidelines for computerized adaptive test development and use in education*. Washington, DC: American Council on Education.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Bhola, D.S., Impara, J.C., & Buckendahl, C.W. (2003). Aligning tests with states' content standards: Methods and issues. *Educational Measurement: Issues and Practice, 22*(3), 21-29.

Hendrickson, A. (2007). An NCME instructional module on multistage testing. *Educational Measurement: Issues and Practice, 26*(2), 44-52.

Huff, K. L. & Sireci, S.G. (2005). Validity issues in computer-based testing. *Educational Measurement: Issues and Practice, 20*(3), 16-25.

Kaira, L. T., & Sireci, S. G. (2010). Evaluating content validity in multistage adaptive testing. *CLEAR Exam Review, 21(2)*, 15-23.

Luecht, R.M. (2005). Some useful cost-benefit criteria for evaluating computer-based test delivery models and systems. *Journal of Applied Testing Technology, 7*(2).

Luecht, R.M. Brumfield, T., & Breithaupt, K. (2006). A testlet assembly design for adaptive multistage tests. *Applied Measurement in Education, 19*(3), 189-202.

Luecht, R. M., de Champlain, A., & Nungester, R. J. (1998). Maintaining content validity in computerized adaptive testing. *Advances in Health Sciences Education, 3*, 29-41.

Martone, A., & Sireci, S. G. (2009). Evaluating alignment among curriculum, assessments, and instruction, *Review of Educational Research 4*, 1332-1361.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational Measurement*, (3rd ed., pp. 13-103). Washington, D.C.: American Council on Education.

Reese, L.M., Schnipke, D.L., & Luebke, S.W. (1999). *Incorporating content constraints into a multi-stage adaptive testlet design* (LSAC Computerized Test Report 97-02). Newtown, PA: Law School Admission Council.

Sireci, S. G. (1998). The construct of content validity. *Social Indicators Research, 45*, 83-117.

Sireci, S.G., Baldwin, P., Martone, A., Zenisky, A.L., Kaira, L., Lam, W., Shea, C.L., Han, K.T., Deng, N., Delton, J., & Hambleton, R.K. (2008). Massachusetts Adult Proficiency Tests Technical Manual, Version 2. *Center for Educational Assessment Research Report No. 677*. Amherst, MA: Center for Educational Assessment, University of Massachusetts.

Thissen, D. & Mislevy, R.J. (1990). Testing algorithms. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (pp. 103-136). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

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.

Zenisky, A. L., Sireci, S. G., Ryan, A., Baldwin, P., Colvin, K., Li, X., Semerjian, A., & Hambleton, R. K. (2012). *The Massachusetts Adult Proficiency Tests technical manual: Version 3*. Amherst, MA: Center for Educational Assessment, University of Massachusetts Amherst.

Table 1

MST Paths Examined in this Study

Path								
Stage1	Stage2	Stage3	Stage4	Stage5	Stage6	Freq.	%	Cum. %
3	3	3	3	3	3	660	7.7	7.7
2	2	2	2	2	2	571	6.7	14.4
1	1	1	1	1	1	512	6.0	20.4
4	4	4	4	4	4	478	5.6	26.0
2	3	3	3	3	3	380	4.4	30.4
3	4	4	4	4	4	372	4.3	34.7
4	3	3	3	3	3	240	2.8	37.5
3	4	3	3	3	3	198	2.3	39.9
2	1	1	1	1	1	185	2.2	42.0
2	3	2	2	3	2	183	2.1	44.2
2	3	2	2	2	2	181	2.1	46.3
2	2	2	2	3	2	168	2.0	48.2
2	4	4	4	4	4	134	1.6	49.8
4	4	3	3	3	3	133	1.6	51.4

Note. The number under each Stage indicates the Educational Functioning Level (1-4) at that particular stage. Thus, the first column (3-3-3-3-3-3) indicates a path where a student started at level 3 and was routed to that level for all subsequent stages.

Table 2

Classification of Path Types

Path						
Stage1	Stage2	Stage3	Stage4	Stage5	Stage6	Path Type
3	3	3	3	3	3	Straight
2	2	2	2	2	2	Straight
1	1	1	1	1	1	Straight
4	4	4	4	4	4	Straight
2	3	3	3	3	3	Increasing
3	4	4	4	4	4	Increasing
4	3	3	3	3	3	Decreasing
3	4	3	3	3	3	Increasing-Decreasing
2	1	1	1	1	1	Decreasing
2	3	2	2	3	2	Increasing-Decreasing
2	3	2	2	2	2	Increasing-Decreasing
2	2	2	2	3	2	Increasing-Decreasing
2	4	4	4	4	4	Increasing
4	4	3	3	3	3	Decreasing

Note. The number under each Stage indicates the Educational Functioning Level (1-4) at that particular stage. Thus, the first column (3-3-3-3-3-3) indicates a path where a student started at level 3 and was routed to that level for all subsequent stages.

Table 3

Descriptive Statistics for Content Validity Indices for Most Common Paths

	Panel					
	A			B		
Path	*n*	*Mean*[a]	*SD*	*n*	*Mean*[a]	*SD*
3-3-3-3-3-3	34	5.08	0.90	38	5.12	1.18
2-2-2-2-2-2	36	5.40	0.83	36	5.01	1.48
1-1-1-1-1-1	34	4.50	1.86	33	4.85	1.47
4-4-4-4-4-4	37	4.84	1.50	37	4.89	1.37
2-3-3-3-3-3	37	5.18	0.88	37	5.11	1.11
3-4-4-4-4-4	35	4.82	1.39	38	5.14	1.37
4-3-3-3-3-3	36	5.08	1.09	37	4.87	1.17
3-4-3-3-3-3	34	5.11	0.90	38	5.08	1.34
2-1-1-1-1-1	34	4.20	1.57	35	4.96	1.43
2-3-2-2-3-2	39	5.24	0.84	35	5.18	1.10
2-3-2-2-2-2	37	5.34	0.82	36	4.93	1.49
2-2-2-2-3-2	38	5.29	0.86	35	5.27	1.07
2-4-4-4-4-4	38	4.94	1.37	37	5.14	1.32
4-4-3-3-3-3	36	5.11	1.09	37	4.83	1.32
Total		**5.01**	**1.14**		**5.03**	**1.30**

Note. “3-3-3-3-3-3” indicates the examinee began in test at level 3 and remained at level 3 for the subsequent five stages.
[a]Ratings are on a 1-6 scale, with 6 indicating that the item measures its intended benchmark “very well.”

Table 4

ANOVA Summary Tables

ANOVA	Source	Sum of Squares	df	Mean Square	F	*p*	_^{2}
Straight Paths by Panel	Path	11.68	3	3.90	2.11	0.10	0.002
	Panel	0.02	1	0.02	0.01	0.91	0.000
	Interaction	4.85	3	1.62	0.88	0.45	0.001
	Error	511.83	277	1.85			
	Total	7555.52	285				
14 Most Popular Paths by Panel	Path	25.02	13	1.93	1.24	0.24	0.001
	Panel	0.09	1	0.09	0.06	0.81	0.000
	Interaction	14.02	13	1.08	0.70	0.77	0.001
	Error	1527.80	986	1.55			
	Total	27333	1014				
One-Way Path Type	Between	10.14	3	3.38	2.19	0.09	0.006
	Within	1557	1010	1.54			
	Total	1567.14	1013				

Table 5

Descriptive Statistics for Content Validity Indices for Path Types

Path Type	*n*	*Mean*	*SD*
Straight	285	4.97	1.36
Increasing	222	5.06	1.25
Decreasing	215	4.93	1.28
Increasing/Decreasing	292	5.18	1.07
		5.04	1.24

Note. Ratings are on a 1-6 scale, with 6 indicating that the item measures its intended benchmark "very well."

Figure 1. MAPT for Reading Multistage-Adaptive Test Design

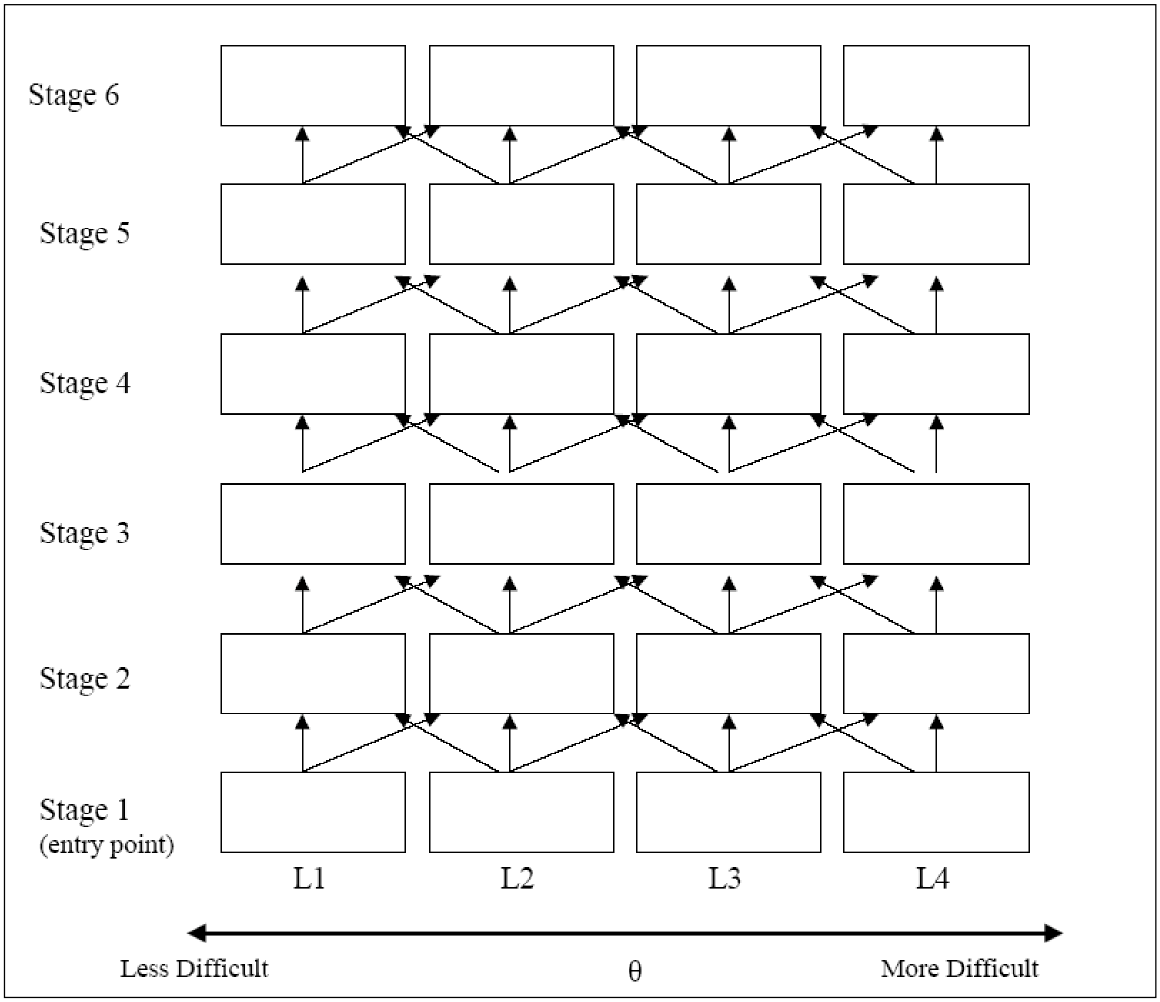

Figure 1. 6-Stage multistage test design for the MAPT. There are four possible entry points and two parallel panels, Panel A and Panel B, (not shown in the Figure), each consisting of four, 40-item tests (one at each level). In this 6-stage design, Stage 1 consists of 15 items, and Stages 2-6 each have 5 items with no items being duplicated across modules or panels. The arrows indicate only some possible paths taken by examinees. All possible paths are not shown.