# *Journal of Computerized Adaptive Testing*

***Volume 1 Number 4***

***September 2013***


# Estimating Measurement Precision in Reduced-Length Multi-Stage Adaptive Testing


**Katrina M. Crotts, April L. Zenisky, Stephen G. Sireci, and Xueming Li**



**DOI 10.7333/1309-0104067**

**The *Journal of Computerized Adaptive Testing* is published by the International Association for Computerized Adaptive Testing**

***www.iacat.org/jcat***

**ISSN: 2165-6592**

International Association for Computerized Adaptive Testing
IACAT
Advancing the Science and Practice of Human Assessment

*Journal of Computerized Adaptive Testing*
*Volume 1, Number 4, September 2013*
DOI 10.7333/1309-0104067
ISSN 2165-6592

# Estimating Measurement Precision in Reduced-Length Multi-Stage Adaptive Testing

**Katrina M. Crotts, April L. Zenisky, Stephen G. Sireci, and Xueming Li**
*University of Massachusetts Amherst*

The extent to which reducing the number of items in a multi-stage adaptive test (MST) affected measurement precision was evaluated. Using the Massachusetts Adult Proficiency Test for Reading (MAPT), a low-stakes MST in adult education, reliability, decision consistency, and decision accuracy estimates were compared for the original and reduced-length tests (from 40 to 35 items). Four different approaches were used: (1) the Spearman-Brown formula, (2) eliminating one item of average discrimination from consecutive stages, (3) completely reassembling new panels, and (4) simulating item responses to the original and shortened MSTs and comparing the standard errors of measurement for simulated examinees. Overall, results suggested comparable levels of measurement precision, improved content representation, and reduced testing time were achievable using the reduced-length tests. The Spearman-Brown estimates were surprisingly close to the estimates based on assembling new panels. Methods for assembling an MST to maintain measurement precision and practical lessons that could generalize to MSTs in other contexts are discussed.

Keywords: *multi-stage adaptive testing, reliability, decision consistency, decision accuracy, response time, test development, validity.*

Multi-stage adaptive tests (MSTs) are becoming increasingly popular due to improved control over test content and other practical features of test administration, while simultaneously being efficient with respect to measurement precision (Luecht, 2005). Unlike item-level computerized-adaptive tests (CATs), MSTs tailor to examinees' proficiency levels by administering a set of items to an examinee, and then determining the next set of items to administer based on cumulative performance on the previous set, rather than on cumulative performance after each item. Similar to a CAT, if the examinee is performing well on the test, a more difficult set of items is administered, and if the examinee is not performing well, an easier set of items is administered (Wainer, 1993).

MSTs offer attractive features over CATs, including control over "test assembly and test form quality control, exposure of test materials, facilitating data management, and reducing re-

quirements for test delivery software to handle complex scoring and item selection algorithms" (Luecht, Brumfield, & Breithaupt, 2006 p.200). In relation to measurement precision, CATs have an advantage over MSTs because the opportunities to adapt to proficiency levels of examinees are related to the number of items on the test. Although MSTs have fewer routing decision points compared to a CAT, which can result in some loss of measurement precision, the loss is typically minimal (Jodoin, Zenisky, & Hambleton, 2006). In this study, the change in measurement precision of the Massachusetts Adult Proficiency Test for Reading (MAPT), a complex MST, was examined when reducing test length.

## Multi-Stage Adaptive Testing Design and Assembly

The design of an MST is defined, in part, by the number of stages within the test. These stages are represented by sets of items called *modules* or *testlets*. An entire set of modules is called a *panel,* which represents all possible sets of items that could be taken by an examinee. For example, if the MST design were 1-3-3, there would be a total of three stages, with the first stage consisting of a single module, and the second and third stages consisting of three modules each. The single module in the first stage is typically referred to as the *routing module* or *locator test* (Hendrickson, 2007), and is typically composed of items that are of moderate difficulty. This module is used to determine which of the three modules will be administered to the examinee in Stage 2, where the modules vary in terms of difficulty—typically easy, medium, and difficult. The examinees' performance on the Stage 2 module is used to update their proficiency estimate to determine the module to which they are routed in Stage 3.

MSTs can vary significantly with respect to design and are not limited to the 1-3-3 design. Luecht and Burgin (2003) identified five different attributes to consider when determining the MST design, including (1) the number of stages; (2) the number of modules per stage; (3) the number of items per stage; (4) the statistical characteristics of the modules within and across stages, including average difficulty and discrimination; and (5) the nature and extent of content, other item attributes, and quantitative test specifications.

Like all test construction endeavors, the goals in developing an MST are maximizing measurement precision (reliability) and meeting content specifications. Ultimately, the goal is an assessment with high measurement precision and content validity. Other considerations include minimal response time in relation to the number of items on the test. Reduced response time can lead to reduced testing costs, especially if the tests are taken in secure locations where the testing organization must pay for seat time.

Developing MST panels can be complex because of constraints and statistical objectives at not only the test level, but also at the module level (Luecht & Burgin, 2003). One common method of developing MSTs is through automated test assembly (ATA). ATA typically uses test information functions (TIFs) as statistical targets and attempts to replicate the intended targets (van der Linden, 1998). TIFs are used to guide the measurement effectiveness of a set of test items along the proficiency continuum. The extent to which target TIFs can be met in operational testing is affected by the quality and number of items on the assessment and in the item bank, as well as the item parameters. Therefore, having a target TIF can help to indirectly control the item difficulties and discriminations for a module (Luecht & Burgin, 2003). It is important to note that although ATA can handle statistical specifications, it does not deal as well with qualitative considerations and aesthetics (Luecht, 2005). Additionally, small item banks can make it more difficult for ATA to be effective. For these reasons, some MSTs, such as the MAPT, are developed

without using ATA to allow for management of reading passages and other content constraints, such as balancing content within modules. In these situations, test developers target the TIF by adding and removing items outside of an ATA algorithm.

## Purpose

This study addressed a very practical issue in MST—whether adequate measurement precision and content domain representation could be achieved using fewer items than were currently being used. That is, can an operational MST be more efficient with respect to item bank usage, content coverage, testing time, and measurement precision? Four methods were explored to determine an answer to this question for the MAPT, a particularly complex MST.

Given a desire to reduce testing time and increase item bank usage, but still maintain adequate content coverage, interest was in predicting the effect on measurement precision if the test length were reduced from 40 items to 35 items. In adult education programs, adult students have limited time to be in a classroom, so assessment time competes with instructional time, and the limited computer resources for taking the exam are in high demand. For most students, the assessment currently takes approximately an hour, so the hope was that 35 items would bring the majority of students' test time to under an hour.

It is well known that all other things being equal, reducing the number of items on a test will reduce score reliability. However, if items can be systematically selected, it is possible to achieve similar, equal, or greater reliability using fewer items. The Spearman-Brown formula (Spearman, 1910) is typically used to answer questions regarding estimates of increased- or reduced-length tests. However, in the case of MSTs, such estimates may be inappropriate due to the content constraints involved in panel assembly. That is, it is not the number of items reduced, but rather *which* items are removed that affect precision and content domain representation.

This study identified and investigated factors that affect measurement precision in an MST, focusing on how a testing program can reduce the length of an MST while maintaining measurement precision and content representation. Specifically, the reliability of a 35-item MAPT compared to the current operational 40-item MAPT was examined using four different approaches, including:

1. The Spearman-Brown formula;
2. Eliminating one item of average discrimination from consecutive stages;
3. Completely reassembling the MAPT using 35 items to make assembled panels with regard to content specifications; and
4. Simulating item responses to the 40- and 35-item MSTs and computing standard errors of measurement for the simulated examinees.

The first strategy was selected because it is easy to calculate and provides a quick estimate. However, given the multiple factors to consider in putting together an MST panel, it is possible that the Spearman-Brown estimate would not be accurate. The second strategy was chosen because by eliminating five items, content factors would make elimination of the lowest discriminating items highly impractical. Eliminating one item of average discrimination from each stage would likely maintain content coverage. Additionally, choosing an average discrimination item was thought to best predict the types of items that would ultimately be removed if the items were carefully selected based on full knowledge of the statistical and content characteristics. The third strategy was just as described above—reassembling the panels based on the entire item bank in a way that would result in the best three panels comprising 35-item tests. This strategy, although

time-consuming, represents the "true" effect on the variables studied, because it reflects what would happen in practice. Reassembling the panels is preferable to the deletion method because with deletion, test information is simply taken away, and the test is confined to items already used. However, when reassembling the panels, the same level of test information as the 40-item test might be able to be reached if the panels are carefully restructured. It is possible to get closer to the desired information level because the full item bank is accessible. Finally, the fourth strategy simulated item responses for both the 40- and 35-item tests to examine standard errors of measurement, which indicates the amount of error around an examinee's score. In addition to evaluating the effect of reducing the number of items on internal consistency reliability, the effects on decision consistency and decision accuracy were also evaluated.

# Method

## Instrument

The MAPT is an MST designed to assess the gains in reading proficiency made by adult education students in Massachusetts. Its content is aligned with the Massachusetts adult education curriculum framework in reading (Adult and Community Learning Services, 2005). The purpose of the MAPT 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 MAPT is a fixed-length MST comprising 40 multiple-choice items. The MST design for the MAPT is presented in Figure 1. What makes the MAPT more complex than most MSTs is that there are five possible *entry points* at which examinees start the test (Levels 1–5; L1–L5). That is, rather than a single point on the proficiency continuum where all examinees start, the MAPT has five starting points that represent different levels of difficulty that correspond to different proficiency levels. This feature is due to the fact that adult learners span a wide educational continuum (similar to grades in K–12 education). Thus, the MAPT scale essentially represents vertical equating of five proficiency level tests[1]. The first time a student takes the MAPT, the classroom teacher assigns the student to an entry point. Thereafter, the most recent MAPT score is used to determine the entry point for all subsequent tests (Sireci et al., 2008). Although these levels are referred to as Levels 1 through 5 here, the National Reporting System (NRS) for Adult Education defines them as Beginning Basic, Low Intermediate, High Intermediate, Low Adult Secondary, and High Adult Secondary.

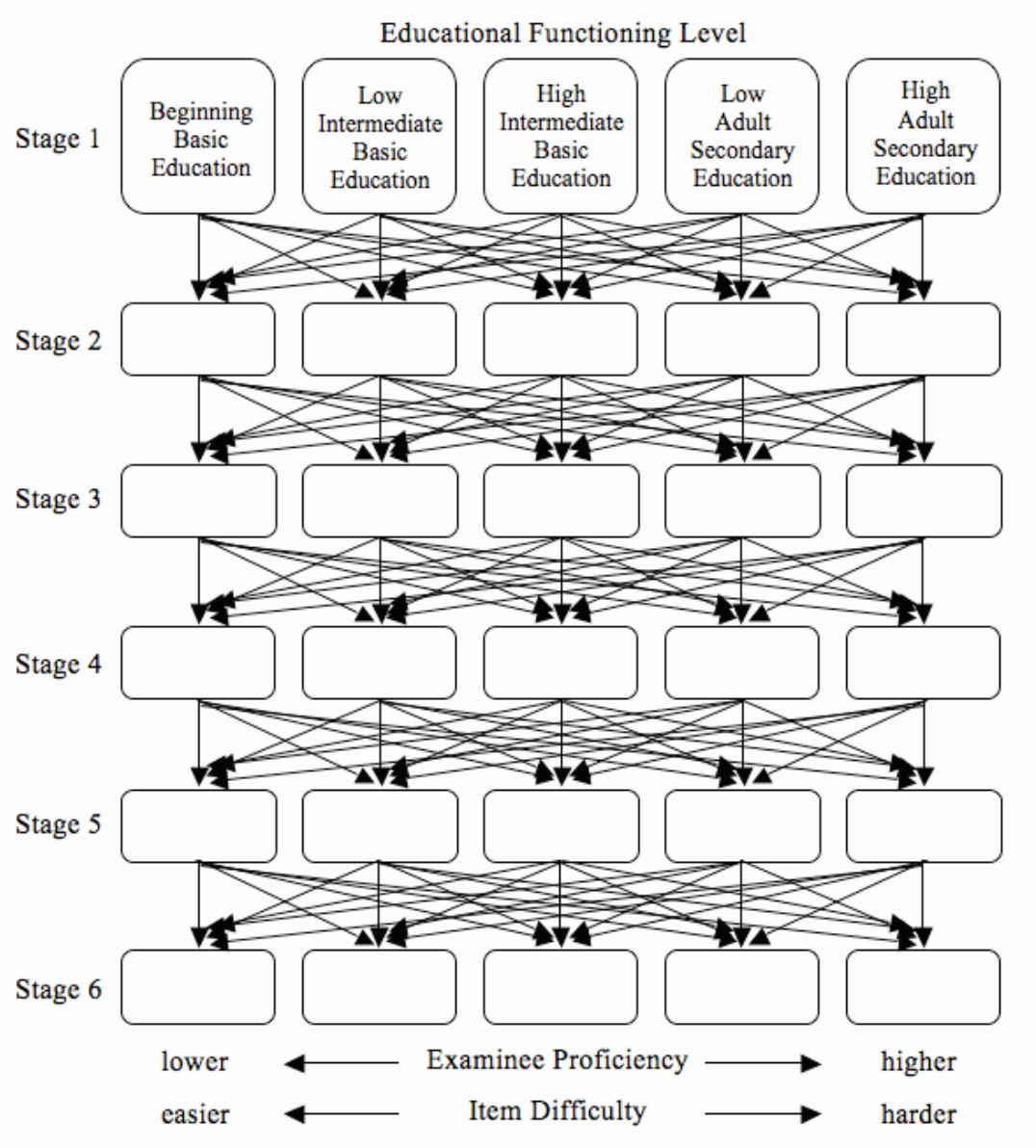
As shown in Figure 1, the MAPT involves six stages. Coupled with the five entry points, there are 30 modules (sets of items) involved in a single panel and three parallel panels (panels A, B, and C), each consisting of five, 40-item tests (one at each level).The first stage consists of 15 items and the remaining stages include five items each; no items are duplicated across modules or panels. Given that 40 items are administered to each examinee, 200 unique items are needed to build just one panel.

To move across stages, an IRT-based estimate of proficiency ($\hat{\theta}$) is calculated at the end of each stage, and this proficiency estimate is used to determine the next module administered (Sireci et al., 2008). This movement through the different stages is called a path, which is essen-

[1] Items were field-tested at each proficiency level and concurrently calibrated onto a common scale. The scaling procedures of the MAPT are beyond the focus of this paper; interested readers are referred to Zenisky, Sireci, Martone, Baldwin, & Lam (2009). The original scaling of the MAPT is described in Sireci et al. (2008).

tially a test "form" for each examinee. The arrows in Figure 1 indicate the possible paths an examinee can take. As shown in Figure 1, it is possible to "travel" to non-adjacent modules. For example, an examinee could start at Level 1 and then be routed to Level 3 in Stage 2. Routing decisions are made using $\hat{\theta}$, which is computed based on all items an examinee answered up to that decision point. For example, at the end of Stage 2 on the 40-item test, a total of 20 items (15 items in Stage 1, five items in Stage 2) would be used to estimate $\theta$. IRT information is maximized for the $\theta$ estimate available at each decision point. Maximization at each decision point is based on computed TIFs for each module within the subsequent stage (Equation 1), and routing the examinee to whichever module yields the most information for their $\theta$ estimate. This entire process is seamless to the examinee (Sireci et al., 2008).

**Figure 1. Six-Stage MAPT MST Design**



Given that assessing reading typically involves passage-based items, assembling a panel for the MAPT can be an arduous task because items based on a reading passage are administered as a set, and the statistics for items within a set can vary enough to limit the extent to which sets can attain target TIFs within and across levels and stages. Currently, the MAPT is administered with three operational panels requiring a total of 600 items, which are needed because students test up to three times per year.

To ensure adequate content representation and alignment to the curriculum frameworks, MAPT panels are assembled as a connection of *straight paths,* which refer to the situation where a student started at a particular test level and remained at that level for the five subsequent stages. Panels are assembled this way because the test specifications are built to represent the content and cognitive targets specified at each level. It is important to note that in assembling multiple panels in MSTs, the best items cannot be selected for the first panel, because then the quality of the other panels would suffer. Thus, all panels must be developed simultaneously.

## Estimating Score Reliability

Score reliability can be estimated from an IRT framework using the formula that relates test information to standard error, and the formula for the standard error of measurement (*SEM*). Test information can be defined as the sum of item information,

$$TIF(\theta) = \sum_{i=1}^{n} \left( \frac{2.89 a_i^2 (1 - c_i)}{\left\{ c_i + \exp\left[ 1.7 a_i (\theta_j - b_i) \right] \right\} \left\{ 1 + \exp\left[ -1.7 a_i (\theta_j - b_i) \right] \right\}^2} \right), \tag{1}$$

where, $a_i$, $b_i$, and $c_i$ are discrimination, difficulty, and guessing parameters for each item based on the 3-parameter logistic model (3PLM),

$$P_i(\theta_j) = c_i + (1 - c_i) \frac{\exp\left[ 1.7 a_i (\theta_j - b_i) \right]}{1 + \exp\left[ 1.7 a_i (\theta_j - b_i) \right]}. \tag{2}$$

Because the MAPT has multiple items associated with a common stem (e.g., a reading passage) local item dependence might have made the 3PLM an inappropriate model for use with the MAPT. However, prior research has shown that even when local item dependence is present, its effects on examinees' $\theta$ estimates are minimal (Zenisky, Hambleton, & Sireci, 2002). The reading passages on the MAPT tend to have only two to four items, which also reduces the likelihood of local item dependence compared to longer item sets (Sireci, Thissen, & Wainer, 1991).

The *SEM* is directly related to test information and is denoted as

$$SEM(\theta) = \frac{1}{\sqrt{TIF(\theta)}}. \tag{3}$$

For the purposes of this study, the focus was on *SEM* at a specific point, specifically where test information was maximized. Focusing on the point on the $\theta$ scale at which test information was maximized makes *SEM* unconditional. Therefore, unconditional *SEM* is also directly related to reliability,

$$SEM = \sigma_x \sqrt{1 - r_{xx'}}, \tag{4}$$

where $\sigma_x$ = the standard deviation of test scores and $r_{xx'}$ = the reliability estimate, an estimate of coefficient alpha that can be obtained by

$$r_{xx'} = 1 - \frac{SEM^2}{\sigma_x^2}. \tag{5}$$

Given the MST design, content specifications, and use of passage-based item sets, it is difficult to estimate what the reliability estimates would be if the length of the MAPT were reduced. To represent the content specifications, it was concluded that at least 35 items would be needed. To evaluate the effect of reducing the fixed length of the MAPT from 40 items to 35 items, reliability was estimated in three different ways: (1) using the Spearman-Brown formula, (2) eliminating one item from Stages 2–6, and (3) completely reassembling three new panels of 35-item tests from the same item bank from which the current 40-item tests were created. To select the items for removal in the second strategy, an item within each of Stages 2–6 was removed that had average item discrimination (which should make the results approximate the Spearman-Brown estimate), while still meeting test content specifications across the entire test.

The Spearman-Brown formula was calculated by

$$r^*_{xx'} = \frac{Kr_{xx'}}{1 + (K-1)r_{xx'}}, \tag{6}$$

where $r^*_{xx'}$ is the predicted reliability, $K$ is the factor by which the length of the test is changed, and $r_{xx'}$ is the reliability of the current test. For the MAPT, the reduction from 40 to 35 items was 12.5%. This means that $K$ would equal .875 when applying the Spearman-Brown formula. In addition, data were also simulated to gauge the standard errors associated with students' $\theta$ estimates in both the 35- and 40-item conditions.

The Spearman-Brown estimate is a theoretical estimate that might be unrealistic in an MST situation. The item-deletion method is a bit more realistic, but not optimal, because the deleted item might not be usable in another panel due to the fact that items might be associated with a common stem (i.e., reading passage, table, or graphic). For example, if the item was associated with a reading passage and deleted, that item would not be usable on another part of that panel because it could lead to a replicated reading passage within a path for an examinee. Therefore, it was concluded that the most appropriate way to estimate the reliability of a 35-item MAPT was to completely assemble three new 35-item panels (i.e., the third strategy above), taking into consideration the content specifications, item sets, item enemies, proficiency level cut-scores that determine where the straight paths should lie to ensure that each level represents the appropriate proficiency associated with that level, and other practical factors. The new 35-item tests contained 15 items in Stage 1 followed by four items in each subsequent stage, compared to five items in each subsequent stage as found in the 40-item test. This 35-item panel assembly reflects the true state of affairs if the test were shortened. To accomplish this task, three new 35-item panels were developed using the same bank of 843 pre-calibrated items that were used to create the panels for the 40-item test.

## 35-Item Panel Construction

The current panel construction process for the 40-item MAPT is completed manually rather than through the use of ATA. Panel construction begins at the easiest level (Level 1) with three panels for each level developed simultaneously to ensure item quality across all panels. For the 35-item assembly, the separation difficulty for each level was determined by the mean and overlap of difficulty parameters for the items within each level and the current 40-item MAPT cut-scores for each level. At the time of the test assembly for both the 40- and 35-item tests, 843 items were available in the item bank. The 40-item tests were assembled and administered operationally prior to this study. For the 35-item tests, panels were assembled using 15 items in Stage 1, followed by four items in each remaining stage (e.g., 15-4-4-4-4-4). With 35 items in each path, five levels, and three panels, a total of 525 items were needed to complete the test construction process (a reduction of 12.5% in items needed). The item bank included information such as the calibrated *a*, *b*, and *c* IRT parameters, item details, and the number of items related to an item passage. Specifically, the item bank had an average *a* parameter of 1.53, an average *b* parameter of 0.64, and an average *c* parameter of 0.21. Within the item bank, *b* parameters ranged from −1.05 to 2.78. When developing panels, the focus was on the straight paths (e.g., staying within one level for all six stages), while attempting to make each module as close to a mini-version of the test specifications as possible.

In developing the MAPT panels, it is first important to consider the test specifications (Table 1) and attempt to match them as closely as possible. As shown in Table 1, items fall within three main categories including: (1) Reading Foundations, (2) Informational Reading, and (3) Literary Reading. In constructing the panels, the goal was to stay within 5% of these specifications. It was also essential to ensure that all panels were unique (i.e., no duplicate items across stages or levels), and that item parameters were balanced across stages within a panel. Because this particular assessment was a reading assessment, it was also necessary to be careful when breaking up items associated with a common stem (i.e., reading passages, tables, or graphics). Specifically, if items with a common stem ranged in difficulty, those items would be split up to target the appropriate difficulty levels. If those items were split, it was essential that the common stem did not appear again in another stage across multiple levels within the same panel. Additionally, the length of the reading passage should also be considered, especially at the lower levels. Thus, panel assembly for the MAPT requires juggling several pieces of the content and statistical puzzle to ensure statistical and content equivalence across panels, and equal separation of difficulty between levels within a panel.

Once the panels were assembled, the module information functions were obtained using a software program specifically designed for MAPT assembly (Baldwin, 2007). Focusing again on the straight paths, TIFs were created for each module (18 in total). TIF criteria for both the 40- and 35-item tests were the same. Essentially, all maximum test information was targeted to be approximately 32, corresponding to a reliability of approximately 0.90, with all TIFs for each level to peak at similar points on the $\theta$ scale. When comparing the TIFs, two main questions were asked: (1) Are intersection points across the observed level-specific TIFs evenly distributed to show the range of $\theta$s? and (2) Are the heights of the TIFs across the levels about the same? If, for some reason, TIFs overlapped, causing one level to not to be seen (i.e., no peak information for that level), adjustments were made. Specifically, the average *b* parameter was increased or decreased within the module to shift the curve left or right, or the average *a* parameter was increased or decreased within the module to shift the curve of the peak. When moving items, it was

essential to maintain the test specifications and average *a* and *b* parameter calculations that were already established.

In addition to identifying the maximum test information for each level, the average level-specific test information across the peak information was calculated. Specifically, the average information was calculated by first identifying the intersecting points along the $\theta$ scale between the level-specific TIFs, essentially the points at which the level-specific TIFs did not overlap with another level-specific TIF. These intersecting points identified the peaks at which average information was calculated. At Levels 1 and 5 (where there were no intersecting points to the left and right of the TIFs, respectively), peak information was identified by taking the average span of $\theta$ across a peaked level-specific TIF. This average span was then subtracted from the Level 1 to Level 2 intersecting point, and added to the Level 4 to Level 5 intersecting point. Once the final 35-item panels were assembled, reliability estimates were calculated using Equation 5. These estimates were then compared to the original 40-item panels, the 35-item panels assembled by deleting one item in Stages 2–6, and the Spearman-Brown calculations.

**Table 1. Test Specifications for the MAPT**

| | Educational Functioning Level | | | | |
|---|---|---|---|---|---|
| Standard | Beginning ABE | Low Intermediate | High Intermediate | Low ASE | High ASE |
| Reading Foundations | | | | | |
| Word ID/Decoding | 15–20% | 5% | 0% | 0% | 0% |
| Vocabulary | 25–30% | 30% | 30% | 25% | 25% |
| Comprehension Strategies | 0–5% | 0–5% | 0–5% | 0–5% | 0–5% |
| Total | 40–45% | 35–40% | 30–35% | 25–30% | 25–30% |
| Informational Reading | | | | | |
| Author Organization & Purpose | 5–10% | 10% | 10% | 10–15% | 10–15% |
| Locating & Using Information & Ideas | 15–20% | 20% | 20% | 20–25% | 20–25% |
| Reliability & Completeness of Information | 0–5% | 5% | 5% | 5% | 5% |
| Synthesis of Ideas | 5% | 5% | 5% | 5% | 5% |
| Total | 35% | 40% | 40% | 40% | 40% |
| Literary Reading | | | | | |
| Literary Structures | 10–15% | 15% | 15% | 15% | 15% |
| Literary Technique/Style | 5% | 5% | 10% | 10% | 10% |
| Making Connections | 5% | 5% | 5% | 5% | 5% |
| Total | 20% | 25% | 30% | 30% | 30% |

## Standard Errors of Measurement

Examining the (conditional) *SEM* associated with each examinee's $\theta$ estimate will allow examination of how closely the score represents the examinee's "true" test score. The *SEM* associ-

ated with an examinee's score indicates the amount of error around the test score, meaning the lower the *SEM*, the closer the test score is to the "true" score. To obtain the *SEM,* dichotomous response data were generated using the 3PLM (Equation 2) for both the 40- and 35-item tests. Generated item response data was based on the "true" item parameters (40-item test item parameters). Similarly, because the uniquely assembled 35-item assessments were not operational, $\theta$s were simulated based on the current distribution of $\hat{\theta}s$ for each straight path. Specifically, the means and standard deviations of $\hat{\theta}$ for Levels 1–5 were –.48(.32), .40(.13), .94(.12), 1.38(.13), and 2.08(.24), respectively. A total of 500 $\theta$s were simulated using R (R Development Core Team, 2011) for each straight path within each of the three panels, for a total of 2,500 $\theta$s per panel. Straight paths were examined due to the fact that a total of 75% of examinees take straight, nearly straight, or mostly straight paths. Specifically, straight paths account for 25% of examinees, nearly straight paths where only one module is different (e.g., 2-3-3-3-3-3) account for 30% of examinees, and mostly straight paths where there are four modules of one level and two of another (e.g., 2-2-2-2-3-3) account for 20% of examinees.

These simulated $\theta$s (2,500 for each panel) and the item parameters based on the current operational item parameters were used to obtain the probabilities to ultimately generate the item response files. Probabilities were based on the 3PLM (Equation 2). These probabilities were then compared to a random number drawn from a uniform distribution. If probabilities were greater than the random number, then an examinee would be assigned a correct response (scored 1), and if the opposite occurred, the examinee would be assigned an incorrect response (scored 0). Using these newly generated responses, an IRT calibration was completed using BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 2003) resulting in an EAP $\theta$ estimate and corresponding standard error. The *SEM* was then calculated based on the calculated test reliability and standard deviation of test score estimates (Equation 4). Finally, the average *SEM*s of the 35-item test were compared to the average *SEM*s of the 40-item test.

### Estimating Response Time

A key reason for reducing the number of items on the MAPT was to shorten the amount of time it takes examinees to complete the test. To estimate the reduction in testing time that would occur if the MAPT were to be shortened to 35 items, the median response times for each item in the current and reduced-length panels were added to estimate a median response time for completing the entire 35-item test. Because the 35-item test was not administered to examinees, but all items had been previously administered as part of the 40-item test, response times were based on the amount of time actual students took to respond to each item as part of the operational 40-item test. Response times from the original 40-item panels were compared to the 35-item panels developed using both the deletion method and the unique assembly method.

## Decision Consistency and Decision Accuracy

The effect of shortening the MAPT on decision consistency and accuracy were of particular interest because the MAPT is a criterion-referenced test that classifies examinees into one of the five Educational Functioning Levels (EFLs) listed in Table 1 (Sireci et al., 2008). $\theta$ cutoff scores between the five levels were used as the criteria to make classifications and included $\theta$s of –.36, .24, .84, and 1.45. Livingston and Lewis (1995) define decision consistency as the classification agreement on two parallel test forms, and decision accuracy as the classification agreement be-

tween the observed scores and true scores, if the true scores could be known.

The software program IRT CLASS (Lee, 2008) was used to calculate both decision consistency and accuracy for each panel. The same simulated $\theta$s used to create the dichotomous response files to obtain the *SEM* were also used here. Specifically, the calculations for the 35-item tests were based on the 2,500 simulated $\theta$s and parameter estimates for the 175 items (525 estimates) for each panel. Results were compared to the results for the operational 40-item panels using the same simulated $\theta$s with a total of 200 items for each panel.

# Results

## Content Representation

The item bank used in developing the MAPT contained 843 items. For the 40-item test, this meant that 71% of the bank had to be used (600 items) and for the 35-item test 62% (525 items) of the bank was used. The 35-item panels were assembled using the same test specifications (Table 1) as the 40-item panels, therefore resulting in comparable content representation across both the 35- and 40-item tests. In fact, all panels in Levels 2, 3, and 4, were within the 5% target of the content specifications. For Levels 1 and 5, however, the 40-item panels differed as much as 10% from the content specifications because of limitations in the item bank. Specifically, at Level 5 more "informational text" was needed, and at Level 1 more "literary" items were needed. This was due to the fact that most bank items were in the difficulty range of $b = 0$ to 1.5. Because of the limited numbers of items at the extreme levels, reducing the number of items within a panel was actually advantageous in staying within 5% of the content specifications.

In addition to representing the content specifications, it is also important to consider the benchmarks (content standards) represented within the assessment. Each of the standards within the content specifications has a subset of benchmarks that are more detailed in explaining the tasks representative of each educational functioning level. It is possible for content specifications to be met, but to only include one benchmark, which would not be representative of that entire content area. For the MAPT, each item is written to a specific benchmark and it is important that each assessment is fairly representative of the benchmarks specific to that level representing reading foundations, informational reading, and literary reading. On the 40-item assessment, approximately 25 benchmarks were represented within each level across three panels on average compared to 22 benchmarks on average for the 35-item assembled assessment. The 40-item test had a range of 22 to 31 benchmarks represented compared to a range of 17 to 27 for the 35-item assembled test. This reduction was expected because fewer items were represented in each of the three standards, thus resulting in fewer benchmarks being represented. Although fewer benchmarks were represented, content specifications were still met for the 35-item panels.

## Test Information and Reliability

Table 2 presents the maximum and average peaked test information for each MAPT Level (entry point) for the current 40-item tests, the 35-item tests based on removing one item from each module at Stages 2–6 (Deleting), and the 35-item tests based on complete reconstruction (Assembly) . Because these results were based on each MAPT Level, they represent an examinee taking a straight path (e.g., 2-2-2-2-2-2), meaning they are routed to the same level upon completion of each module. The maximum test information for each MAPT Level for the 40-item test ranged from 36 to 62 with average peak information ranging from 33.68 to 55.73. Results for the

**Table 2. Maximum, Mean (*M*), and Standard Deviation (*SD*) of MAPT Test Information by Level and Panel**

| No. of Items and Panel | Level 1 | | | Level 2 | | | Level 3 | | | Level 4 | | | Level 5 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Max | *M* | *SD* | Max | *M* | *SD* | Max | *M* | *SD* | Max | *M* | *SD* | Max | *M* | *SD* |
| 40 Items | | | | | | | | | | | | | | | |
| Panel A | 54 | 49.29 | 4.56 | 62 | 55.73 | 6.09 | 45 | 43.71 | 1.50 | 42 | 39.08 | 3.21 | 36 | 33.68 | 2.05 |
| Panel B | 40 | 37.18 | 2.97 | 51 | 46.96 | 3.71 | 46 | 43.71 | 2.60 | 38 | 35.92 | 2.88 | 36 | 34.29 | 1.96 |
| Panel C | 45 | 41.02 | 4.99 | 51 | 48.02 | 2.81 | 49 | 46.09 | 2.60 | 42 | 38.02 | 3.92 | 36 | 34.39 | 1.32 |
| Average | 46.3 | 42.50 | 4.17 | 54.7 | 50.24 | 4.20 | 46.7 | 44.50 | 2.23 | 40.7 | 37.67 | 3.34 | 36.0 | 34.12 | 1.78 |
| 35 Items (Deleting) | | | | | | | | | | | | | | | |
| Panel A | 48 | 43.47 | 4.16 | 56 | 49.60 | 5.49 | 40 | 37.98 | 1.93 | 36 | 33.68 | 2.93 | 31 | 29.38 | 2.00 |
| Panel B | 34 | 30.81 | 3.32 | 45 | 40.37 | 4.10 | 40 | 37.50 | 2.43 | 33 | 30.87 | 2.28 | 31 | 29.12 | 1.97 |
| Panel C | 40 | 36.62 | 3.97 | 44 | 41.69 | 2.46 | 43 | 40.47 | 2.41 | 36 | 33.11 | 3.54 | 30 | 28.93 | 1.34 |
| Average | 40.7 | 36.97 | 3.81 | 48.3 | 43.89 | 4.02 | 41.0 | 38.65 | 2.26 | 35.0 | 32.55 | 2.92 | 30.7 | 29.14 | 1.77 |
| 35 Items (Assembly) | | | | | | | | | | | | | | | |
| Panel A | 34 | 30.54 | 3.52 | 40 | 38.11 | 2.41 | 44 | 40.11 | 3.96 | 33 | 31.04 | 2.28 | 26 | 25.47 | .70 |
| Panel B | 44 | 41.47 | 2.27 | 42 | 40.56 | 1.40 | 41 | 38.18 | 3.11 | 34 | 31.90 | 1.71 | 31 | 29.52 | 1.34 |
| Panel C | 38 | 35.99 | 2.14 | 38 | 36.85 | 1.38 | 34 | 33.02 | 1.18 | 35 | 33.42 | 2.00 | 33 | 30.72 | 1.83 |
| Average | 38.7 | 36.00 | 2.64 | 40.0 | 38.51 | 1.73 | 39.7 | 37.10 | 2.75 | 34.0 | 32.12 | 2.00 | 30.0 | 28.57 | 1.29 |

35-item test by deleting one item had a maximum test information range of 30 to 56 with an average peak test information range of 28.93 to 49.60, and the 35-item test by assembly maximum information ranged from 26 to 44 with an average peak information range of 25.47 to 41.47. It is important to note that the maximum information was not always equal to the average peak information because the maximum information was not always centered within the peaked information. Figure 2 shows the range of test information curves for both the 40-item test and 35-item test by assembly.

Because test information is not a scale-free metric, the reliability estimate for each MAPT level was examined. Table 3 presents the reliability estimates for each MAPT level for the current 40-item tests, Spearman-Brown estimates, the 35-item test developed by selectively removing items from Stages 2–6, and the 35-item test developed by complete reassembly. The estimates for the 40-item test were all above 0.90, which was expected because the maximum test information for each level was above 32. The Spearman-Brown estimates also yielded reliability estimates above 0.90 at all levels except for the highest level (Level 5). As predicted, the 35-item tests developed by eliminating an average discriminating item at each stage (while protecting the test specifications) produced results similar to the Spearman-Brown estimates. The assembled 35-item panels yielded slightly lower, but similar, results to the both the Spearman-Brown estimates and 35-item estimates based on item deletion. All reliability estimates were greater than 0.87.

## Standard Errors of Measurement

The analysis of the *SEM*s based on the simulated student responses for the 40- and 35-item tests indicate the accuracy of student test scores on the original assessment and the newly assembled 35-item assessment. Table 4 summarizes these results. The average *SEM*s across panels were lower for the 40-item test compared to the 35-item test, which was expected since estimates are likely to be more accurate with more test items. Focusing on the average across panels, differences in level-specific *SEM*s between the 40- and 35-item tests were slightly different, but were fairly small and non-consequential, with differences ranging between .01 and .04.

## Estimated Response Time

The testing times estimated from median item response times are reported in Table 5. Testing time for the 40-item test ranged from approximately 44 to 63 minutes depending on the path taken. Results for the 35-item test developed by deletion yielded a reduction of testing time ranging from 5 to 10 minutes, which represented a reduction of testing time between 12% and 16%, on average. Results for the 35-item test developed by panel assembly yielded an even higher reduction of testing time ranging from 7 to 15 minutes for Levels 1–4. However, Level 5 was only reduced by one minute when compared to the 40-item test. For Levels 1–4, this represented a reduction of testing between 12% and 24%, on average.

In addition to a test-level examination of items across the different tests, the proportion of items falling into different response time categories was also examined to determine if the 40-item and 35-item tests had proportionally the same number of items in each response time category, to ensure the accuracy of time estimates. Table 6 shows that the proportions across each form were roughly equivalent.

**Figure 2. TIFs for 40-Item and 35-Item Tests (Assembly)**

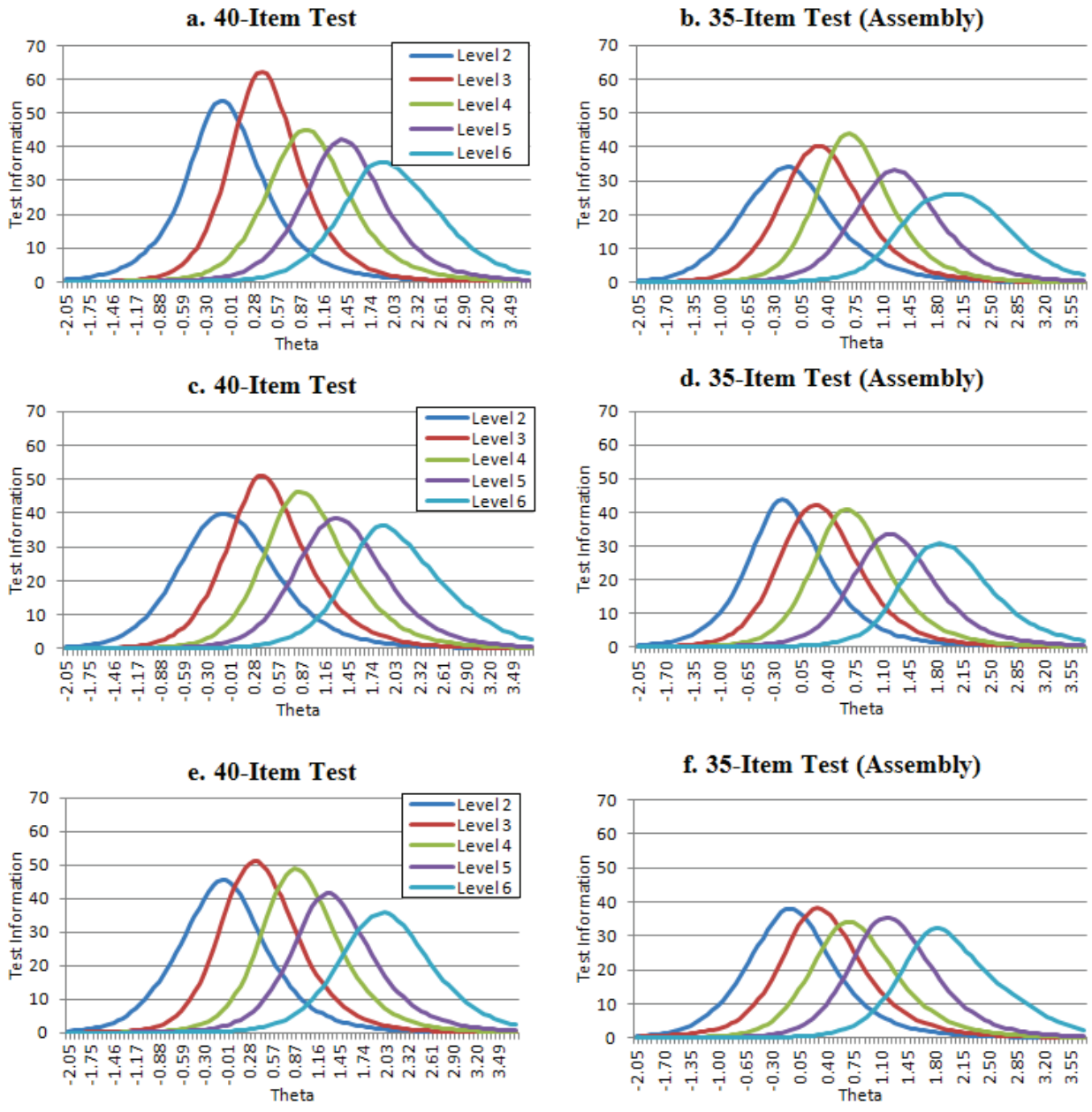


**Table 3. Reliability Estimates by Test Length, Panel, and Level**

| Number of Items and Panel | Level | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 |
| 40 Items | | | | | |
| Panel A | 0.940 | 0.948 | 0.928 | 0.923 | 0.910 |
| Panel B | 0.919 | 0.937 | 0.930 | 0.915 | 0.910 |
| Panel C | 0.928 | 0.937 | 0.934 | 0.923 | 0.910 |
| Spearman-Brown (Reduced to 35 Items) | | | | | |
| Panel A | 0.932 | 0.941 | 0.919 | 0.913 | 0.899 |
| Panel B | 0.909 | 0.928 | 0.921 | 0.904 | 0.899 |
| Panel C | 0.919 | 0.928 | 0.925 | 0.913 | 0.899 |
| 35 Items (Deleting) | | | | | |
| Panel A | 0.933 | 0.942 | 0.919 | 0.910 | 0.896 |
| Panel B | 0.905 | 0.928 | 0.919 | 0.902 | 0.896 |
| Panel C | 0.919 | 0.927 | 0.925 | 0.910 | 0.892 |
| 35 Items (Assembly) | | | | | |
| Panel A | 0.905 | 0.919 | 0.927 | 0.902 | 0.876 |
| Panel B | 0.927 | 0.923 | 0.921 | 0.905 | 0.896 |
| Panel C | 0.915 | 0.915 | 0.905 | 0.906 | 0.902 |

**Table 4. Average Standard Errors of Measurement Across Tests and Panels**

| Number of Items and Panel | Level | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 |
| 40 Items | | | | | |
| Panel A | 0.161 | 0.173 | 0.209 | 0.210 | 0.214 |
| Panel B | 0.198 | 0.196 | 0.201 | 0.222 | 0.215 |
| Panel C | 0.184 | 0.192 | 0.197 | 0.208 | 0.218 |
| Average | 0.181 | 0.187 | 0.202 | 0.214 | 0.216 |
| 35 Items (Assembly) | | | | | |
| Panel A | 0.210 | 0.220 | 0.206 | 0.245 | 0.258 |
| Panel B | 0.177 | 0.214 | 0.214 | 0.236 | 0.227 |
| Panel C | 0.199 | 0.232 | 0.238 | 0.240 | 0.225 |
| Average | 0.195 | 0.222 | 0.219 | 0.240 | 0.237 |

**Table 5. Median Response Time (In Minutes) by Level and Panel**

| Number of Items and Panel | Level 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 40 Items | | | | | |
| Panel A | 43.4 | 51.6 | 65.6 | 45.3 | 48.1 |
| Panel B | 43.4 | 46.2 | 62.7 | 59.0 | 45.8 |
| Panel C | 44.7 | 58.6 | 59.8 | 62.7 | 44.5 |
| Average | 43.8 | 52.1 | 62.7 | 55.6 | 46.1 |
| 35 Items (Deleting) | | | | | |
| Panel A | 38.6 | 42.7 | 59.4 | 37.0 | 39.0 |
| Panel B | 36.8 | 42.1 | 47.0 | 52.8 | 40.3 |
| Panel C | 39.7 | 52.9 | 52.3 | 54.3 | 39.6 |
| Average | 38.4 | 45.9 | 52.9 | 48.1 | 39.6 |
| 35 Items (Assembly) | | | | | |
| Panel A | 30.4 | 44.9 | 44.2 | 49.9 | 44.0 |
| Panel B | 31.5 | 48.7 | 55.1 | 46.8 | 38.6 |
| Panel C | 37.9 | 42.6 | 45.3 | 49.3 | 53.5 |
| Average | 33.3 | 45.4 | 48.2 | 48.7 | 45.4 |

**Table 6. Proportion of Items in Each Response Time Category**

| Panel and No. of Items | Median Response Time (In Seconds) ≤ 20 | 21–40 | 41–60 | 61–80 | 81–100 | 101–120 | 121–140 | ≥ 141 |
|---|---|---|---|---|---|---|---|---|
| Panel A | | | | | | | | |
| 40 Items | 1.0 | 21.5 | 29.0 | 20.0 | 7.0 | 5.5 | 4.5 | 11.5 |
| 35 Items (Deleting) | 1.1 | 21.7 | 29.1 | 20.0 | 7.4 | 5.1 | 4.0 | 11.4 |
| 35 Items (Assembly) | 0.6 | 23.4 | 29.1 | 16.6 | 10.3 | 8.0 | 2.9 | 9.1 |
| Panel B | | | | | | | | |
| 40 Items | 0.0 | 18.0 | 32.5 | 17.0 | 11.5 | 10.5 | 2.0 | 8.5 |
| 35 Items (Deleting) | 0.0 | 19.4 | 31.4 | 17.1 | 12.0 | 9.7 | 2.3 | 8.0 |
| 35 Items (Assembly) | 0.6 | 17.1 | 29.7 | 20.6 | 13.7 | 6.9 | 3.4 | 8.0 |
| Panel C | | | | | | | | |
| 40 Items | 0.0 | 17.0 | 32.0 | 15.0 | 13.0 | 7.0 | 3.5 | 12.5 |
| 35 Items (Deleting) | 0.0 | 17.1 | 32.0 | 15.4 | 12.0 | 6.9 | 4.0 | 12.6 |
| 35 Items (Assembly) | 0.6 | 22.3 | 29.1 | 17.7 | 6.9 | 8.0 | 5.7 | 9.7 |

## Decision Consistency and Accuracy

Results for decision consistency and accuracy are shown in Table 7. To obtain decision consistency, the proportion of learners consistently classified into the same EFL was summed. Results across all three 40-item panels indicated that approximately 81% of examinees were consistently classified into the same EFL. This result indicates the percentage of learners that would be classified into the same EFL if they immediately took a parallel form of the test. For decision accuracy, results for all three 40-item panels indicated that about 86% of examinees were accurately classified into their "true" EFL. Both false positives (examinees misplaced into higher EFLs) and false negatives (examinees misplaced into lower EFLs) were relatively small, between 6% and 8%.

For the 35-item test by assembly, slightly lower decision consistency results were found for all three panels with approximately 78–79% of examinees consistently classified into the same EFL. Similarly, decision accuracy results were also slightly lower with slightly higher false positive and false negative rates. For all three 35-item assembled panels, approximately 85% of examinees were accurately classified into their "true" EFL. Both false positive and false negatives were between 7% and 8%. Although there was a 1% to 1.5% difference on the false positives and negatives between the 40- and 35-item test panels, these differences are not consequential given the low-stakes nature of the MAPT at the student level. However, this result could be a concern for ABE programs where a student-level gain is important for receiving funding. Since 9,113 took the 40-item assessment, a 1% to 1.5% increase in false positives and negatives would result in approximately 91 to 137 students being falsely categorized.

**Table 7. Decision Consistency and Accuracy, and False Positives and Negatives, by Test Length and Panel**

| Number of Items and Panel | Decision | | False | |
|---|---|---|---|---|
| | Consistency | Accuracy | Positives | Negatives |
| 40 Items | | | | |
| Panel A | 0.810 | 0.862 | 0.070 | 0.068 |
| Panel B | 0.809 | 0.861 | 0.064 | 0.075 |
| Panel C | 0.811 | 0.863 | 0.067 | 0.070 |
| 35 Items (Assembly) | | | | |
| Panel A | 0.783 | 0.846 | 0.079 | 0.075 |
| Panel B | 0.786 | 0.850 | 0.073 | 0.079 |
| Panel C | 0.782 | 0.845 | 0.079 | 0.076 |

## Discussion and Conclusions

This study was motivated by practical concerns related to efficient use of an MST item bank, a motivation to reduce testing time, and a desire to maintain measurement precision and content coverage. In general, the results suggested substantial time savings might occur by reducing the number of items on this MST without sacrificing adherence to the content specifications and with a non-consequential loss of measurement precision. If the reduced-length assessment is assembled through careful test construction efforts that consider content specifications, item sets (e.g., items associated with a reading passage), item enemies, and proficiency level cut-scores,

reliability estimates, standard errors of measurement, decision accuracy and consistency estimates, and adherence to the content specifications can all be roughly comparable to the original, longer-length test.

The newly constructed testing panels represented the panels that could be developed under the 35-item constraint. Results showed that the test could in fact be reduced to 35 items while retaining comparable measurement precision. These panels represented the criteria for actual measurement properties, but it was interesting that the Spearman-Brown estimate, and the systematic deletion of items, provided accurate estimates of the optimal (assembly) solution. This is encouraging for future research where a quick estimate is needed, without having to go through the trouble of assembling entire new panels. However, the generalizability of this result is probably dependent on the quality of items in the bank. The 843 items from which the operational and new panels were constructed represented items that were previously field-tested, reviewed by content experts, and deemed appropriate for operational use.

A key piece in obtaining adequate measurement precision for the 35-item assessment was through careful test development using IRT. The task of assembling test panels must be done carefully, considering factors such as content specifications, item sets, item enemies, proficiency level cut-scores that determine where the straight paths should lie, balancing of content within modules, management of reading passages with respect to passage length and potential overlap across levels and panels, balancing of item parameters across all modules, and close examination of test information functions for each stage and for the full test. This thorough process helped lead to the desired result of strong measurement precision throughout the straight paths for the MAPT. Additionally, the advantage of developing a 35-item test meant that items were more carefully selected because more bank items were available. For this reason, the shorter 35-item tests had comparable decision accuracy and consistency results to the 40-item tests.

## Practical Lessons

Although the generalizability of this study is limited due to the fact that the study examined only one MST, the issues pertaining to measurement precision probably generalize to MSTs in other contexts. In determining whether to reduce the number of items on an MST, it is first important to carefully evaluate the current measurement precision. This is difficult to do in the case of an MST because of the number of possible routes an examinee could take depending on both his or her proficiency level and the complexity of the MST design. In this particular context, the MST design was fairly complex with five levels (entry points), six stages within each path with 15 items in Stage 1 followed by five items in the following five stages (reduced to four items in the 35-item test), and three parallel panels. To evaluate measurement precision on the original 40-item test, reliability estimates at each level across the three panels were examined focusing on only the straight paths. Because all paths were not considered, the aim was a coefficient alpha around 0.90 or higher.

Fortunately, because the 40-item test scores exhibited adequate reliabilities, and the Spearman-Brown estimates indicated that high reliability could be obtained with a reduced test of 35-items, the study was continued. In the current context, the Spearman-Brown estimates were thought to produce upper-bound estimates, and so it was important to completely reconstruct the test using optimal 35-item panels rather than relying strictly on the Spearman-Brown estimates. However, had the Spearman-Brown estimates yielded significantly lower reliability with 35 items, it would most likely indicate that a reduction in test length would be problematic. Therefore, when considering reducing the test length of an MST, it is helpful to first examine the

Spearman-Brown estimates to estimate the likely upper bound of test score reliability.

Some practical lessons in examining the possibility of reducing the test length of an MST are to first understand the MST design and item bank. Knowing where in the MST design to reduce the length is an important consideration for balancing content and for retaining accurate proficiency estimation. In the case of the MAPT, the development of the four-item modules was difficult due to the use of reading passages; therefore, future research will investigate removing five items from Stage 1 instead of removing one item from Stages 2–6.

Understanding the item bank is also important for developing panels. If the test bank has a significant number of high- or low-discriminating items that are typically hidden within a module, it might be difficult to reduce the number of items on the test while maintaining strong measurement precision. Knowing the range and distribution of item parameters will assist in determining whether test length should be reduced. Even in MSTs using automated test assembly, knowing where in the MST design to reduce length and knowing the capacity of the prospective item bank are still important factors to consider. Additionally, knowing how to carefully assemble the test or what constraints to consider in the automated assembly will also impact the optimal test design.

## Limitations

One major limitation to this study is its generalizability to other MSTs. The results of this study were strictly related to the test construction process of the MAPT, the item bank, reading content, and MST design. However, important practical lessons related to MST test development and measurement precision are likely to be generalizable. Additionally, although item responses were generated to investigate standard errors, the 35-item test was never administered to real examinees.

## Implications for the MAPT and Future Research

Because this assessment was a reading assessment, there was an additional constraint to consider when developing test panels. Items associated with one specific reading passage had to be considered while attempting to retain content balancing within four-item modules. Because reading passages are either informational or literary and typically contain three or more items, it can be difficult to balance content within a four-item panel. Many modules had reading foundations and informational or literary text, but not all three. Therefore, in relation to the MAPT, it might be better that future 35-item panels be assembled using ten items in Stage 1 followed by five items in the modules constructed for Stages 2–6. Because routing decisions should be made using a range of content whenever possible, the five-item modules would aid in retaining that range of content. Additionally, from a content perspective, using five-item modules will help maintain representation of the content specifications while retaining measurement precision.

Future research in this area should consider constructing the shorter tests and then actually administering them to examinees, resulting in operational item responses rather than simulated item responses. For future research, it is recommended that decision consistency and accuracy be recalculated with IRT CLASS (Lee, 2008) using the operational item responses rather than simulated item responses. Future research should also consider investigating test information and measurement precision of other paths to which examinees were routed. Lastly, future research regarding measurement precision and MSTs in general should consider how different MST designs, as well as different combinations of item numbers within modules, might affect measurement precision.

## Conclusions

In this study, the question of whether comparable measurement precision could be achieved on an MST assessment using fewer items was addressed. Using the MAPT results indicated that through careful test assembly and consideration of test information across the MST design, measurement precision with a 35-item test was fairly comparable to that attained by the original 40-item test. Results also suggested that testing time would be noticeably reduced when using the reduced-length test. Decision consistency and accuracy results produced slightly lower classification results for all three 35-item assembled panels compared to the 40-item panels. The results are encouraging—they indicate that if a test developer is strategic in selecting fewer items from an operational MST, it might be possible to increase the efficiency of an MST without serious reductions in measurement precision.

## References

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

Baldwin, P. (2007). OWL FILE CREATOR [Computer program]. Amherst, MA: School of Education Center for Educational Assessment.

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

Jodoin, M. G., Zenisky, A. L., & Hambleton, R. K. (2006). Comparison of the psychometric properties of several computer-based test designs for credentialing exams with multiple purposes. *Applied Measurement in Education*, 19, 203-220. CrossRef

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement*, *32*, 179-197. CrossRef

Lee, W. (2008). *Classification consistency and accuracy for complex assessments using item response theory* (CASMA Research Report No. 27). Iowa City, IA: University of Iowa.

Luecht, R. M. (2005). Some useful cost-benefit criteria for evaluating computer-based test delivery models and systems. *Journal of Applied Testing Technology, 7*. Retrieved from http://data.memberclicks.com/site/atpu/Volum%207%20Some%20useful%20cost%20benefit.pdf

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

Luecht, R.M., & Burgin, W. (2003, April). *Test information targeting strategies for adaptive multistage testing designs*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

R Development Core Team. (2011). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.RProject.org

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.

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement, 28*(3), 237-247. CrossRef

Spearman, C. C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 3*, 271–295

van der Linden, W. J. (1998). Optimal assembly of psychological and educational tests. *Applied Psychological Measurement, 22*, 195-211. CrossRef

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. CrossRef

Zenisky, A. L., Hambleton, R. K., & Sireci, S. G. (2002). Identification and evaluation of local item dependencies in the Medical College Admissions Test. Journal of Educational Measurement, 39, 291-309. CrossRef

Zenisky, A. L., Sireci, S. G., Martone, A., Baldwin, P., & Lam, W. (2009). Massachusetts adult proficiency test technical manual supplement: 2008-2009. *Center for Educational Assessment Research Report No. 715*. Amherst, MA: Center for Educational Assessment, University of Massachusetts.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (2003). BILOG-MG 3 for Windows: Multiple-group IRT analysis and test maintenance for binary items [Computer program]. Skokie, IL: Scientific Software International, Inc.

## Acknowledgments

This research was funded, in part, through a contract with the Massachusetts Department of Elementary and Secondary Education (Contract Number S1390000000048). The authors appreciate this support. 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.

## Author Address

Katrina Crotts, Educational Testing Service, 660 Rosedale Road, MS-10R, Princeton, NJ 08541. Email: Kcrotts@ets.org