J·T·L·A

The Journal of Technology, Learning, and Assessment

Volume 8, Number 3 · January 2010

Controlling Test Overlap Rate in Automated Assembly of Multiple Equivalent Test Forms

Chuan-Ju Lin

www.jtla.org

A publication of the Technology and Assessment Study Collaborative
Caroline A. & Peter S. Lynch School of Education, Boston College

J·T·L·A

Volume 8, Number 3

Controlling Test Overlap Rate in Automated Assembly of Multiple Equivalent Test Forms

Chuan-Ju Lin

Editor: Michael Russell
russelmh@bc.edu
Technology and Assessment Study Collaborative
Lynch School of Education, Boston College
Chestnut Hill, MA 02467

Copy Editor: Jennifer Higgins
Design: Thomas Hoffmann
Layout: Aimee Levy

JTLA is a free online journal, published by the Technology and Assessment Study Collaborative, Caroline A. & Peter S. Lynch School of Education, Boston College.

Preferred citation:

Lin, C-J. (2010). Controlling Test Overlap Rate in Automated Assembly of Multiple Equivalent Test Forms. *Journal of Technology, Learning, and Assessment, 8*(3). Retrieved [date] from http://www.jtla.org.

J·T·L·A

Abstract:

Assembling equivalent test forms with minimal test overlap across forms is important in ensuring test security. Chen and Lei (2009) suggested an exposure control technique to control test overlap-ordered item pooling on the fly based on the notion that test overlap rate – ordered item pooling for the first t examinees is a function of test overlap rate – ordered item pooling for the previous (t-1) examinees. The exposure control procedure to manage test overlap-ordered item pooling on the fly appears to meet the needs of controlling test overlap rate for tests assembled sequentially. To develop a better understanding of how well the ordered-item-pooling control method functions in automated assembly of multiple forms with the weighted deviations model (WDM) heuristic, this study evaluated its performance under different conditions of test length and test-content specifications by comparing the outcomes to those from the corresponding baseline automated-test-assembly (ATA) conditions, where test overlap controls were not considered. The evaluation criteria included (i) the conformity to the test-assembly constraints, (ii) test parallelism in terms of the resultant psychometric properties, (iii) average test overlap rate, and (iv) distribution of item exposure rate. The results showed that the ordered-item-pooling control procedure demonstrated its effectiveness in most experimental conditions by achieving an acceptable average test overlap rate across multiple forms without compromising the conformity to the test-assembly constraints and the test equity of the assembled forms. Moreover, test security might be ensured in less supportive contexts for ATA by imposing item exposure control together with test overlap control that would be less likely to compromise test quality. More research is needed to verify this hypothesis.

J·T·L·A

Controlling Test Overlap Rate in Automated Assembly of Multiple Equivalent Test Forms

Chuan-Ju Lin
National University of Tainan, Taiwan

Introduction

The automated assembly of multiple test forms for online delivery offers an alternative to a single, computer-administered, fixed test form or even a computerized adaptive test. The constructed forms are usually assembled according to content and psychometric specifications obtained from a reference test. A reference test refers to a test form that has been administered previously and has exhibited acceptable results in terms of form difficulty, variability, passing rate, or other psychometric considerations. If the assembled tests all meet these reference specifications, the test forms can be regarded as equivalent in some sense. For example, if the psychometric specifications refer to a target test information function (TIF) or target test characteristic function (TCF), then the assembled test forms would be TIF or TCF parallel if all of the psychometric specifications were met for all test forms (McDonald, 1999, p. 351). Accordingly, equivalence would in fact mean parallelism. The result is that a single passing standard or score could be used across forms without post-administration equating or the establishment of separate passing scores for each form.

Theoretically, if pre-equated test forms are truly parallel, post-test equating or the establishment of separate passing scores for each form may not be required. The reason for this is because the differences in a candidate's scores on multiple test forms should occur from random fluctuation instead of systematic differences in the test forms. Accordingly, assembling multiple test forms before administration is an appealing idea because pre-equated (parallel) test forms could be administered to examinees in high-stakes testing situations with a nominal amount of post-administration delay in reporting scores when designed properly.

However, the multiple equivalent forms may or may not consist of unique test items. In automated assembly of multiple equivalent test forms, a test assembly algorithm selects test items according to their

ability to satisfy a set of particular constraints. Items with greater ability to fulfill all constraints may be presented in most forms, whereas items with less ability to fulfill all constraints may never be used. Additionally, when item pools from which the forms are assembled are small relative to the required length of the test forms and the number of forms required, individual items may have to appear on more than one form. If enough items appear frequently on many forms, test security and validity will be in question, and the cost of developing and maintaining item pools will increase.

Therefore, test overlap control is an important concern, and may be more crucial for automated test assembly (i.e., ATA) than for computerized adaptive testing (i.e., CAT). In CAT, it is well recognized that the test overlap rate tends to be much greater for examinees of similar ability. Similarly, in the context of assembling equivalent test forms, the test overlap rate would be extremely high because the items selected to fulfill the constraints (e.g., target test information function) would be almost the same across multiple test forms without exposure control.

Accordingly, one of the goals of the test assembly process should be to minimize test overlap rate, defined as the percentage of items shared between any two forms. In automated test assembly, one way to do this is to include item usage as another constraint or target in the solution of the assembly problem. However, this may be unnecessary, especially if the test-assembly process is burdened with numerous other constraints such as multiple levels of content classifications and key balancing requirements, in addition to the psychometric requirements of the tests. That is, any constraint that forces items onto a test form may end up doing so at the expense of other constraint goals. To achieve an acceptable average test overlap rate (ATOR) across multiple forms without compromising the conformity to the test-assembly constraints, it may be more efficient to implement item exposure control outside the solution of the assembly process. Research with more comprehensive investigation of such an approach to effectively control the test overlap rate in automated test assembly may be useful.

Traditional Procedure of Item Exposure Control in Computer Based Testing

The typical procedure used to control for item exposure in CAT is a conditional approach first proposed by Sympson and Hetter (1985). The Sympson and Hetter (SH) control algorithm is based on the concept of conditional probability: $P_i(S,A) = P_i(A \mid S) \times P_i(S)$, where $P_i(S,A)$ is the probability that item i is selected and administered for a testing administration, $P_i(S)$ is the probability that item i is selected, and $P_i(A \mid S)$ is the proba-

bility of administering an item, given that it has been selected. $P_i(A|S)$ is also regarded as an item's exposure control parameter. Given a maximum expected item-exposure rate or a target-exposure rate (r), if $P_i(S,A)$ is replaced by r, the goal of the SH probabilistic method is to obtain the item exposure control parameters for all items in the pool with an iterative simulation process. The item exposure control parameters control the item administration such that every item is administered no more than $r100\%$ of the time, where $0 \leq r \leq 1$.

Exposure Control at the Item and the Test Levels

It is important to track item exposure at the item level and at the test level by monitoring item exposure rate and test overlap rate. Item exposure rate is defined as the percentage of all exams in which an item is administered. Test overlap rate is usually defined as the percentage of items shared by a pair of exams (or tests) of a fixed length. However, most research has paid more attention to item exposure control at the item level than at the test level. With high test overlap rates across multiple equivalent forms, test takers are very likely to obtain test information from those who have taken an alternative form. To ensure test security, controlling test overlap rate is necessary in the context of continuously administering alternate test forms.

To investigate how to simultaneously control item exposure at the item and test levels in CAT, Chen, Ankenmann, and Spray (2003) derived an algebraic function to reveal the relationship between the average test overlap and item exposure rate. Specifically, with a large-sample approximation,

$$\overline{T} = \frac{S^2 + \mu^2}{\mu}$$

where $\overline{T}$ is the average test overlap rate between pairs of CATs, and μ and S^2 are the mean and variance of item exposure rates, respectively. Accordingly, the test overlap rate can be governed by controlling the mean and variance of the item exposure rates. Additionally, they suggested that the value of μ could be considered to be a maximum allowable rate for any single item, and thereby item exposure control methods which require a specification of maximum item exposure rate (e.g., Sympson & Hetter, 1985) would yield the most direct control at the test level as well as at the item level. Following Chen et al.'s suggestion of controlling the mean and variance of the item exposure rates, Chen and Lei (2005) developed an item exposure control method to provide item exposure control at both

the item and test levels. The method proposed by Chen and Lei (2005) is an extension of the Sympson and Hetter (SH) procedure, and is called the SH procedure with test overlap control (SHT).

However, acknowledging that stable item exposure parameters need to be derived from time-consuming iterative simulations in SH and SHT procedures before operational test administration, recent research shows interest in controlling item exposure on the fly or during test administration. For example, Chen, Lei, and Liao (2008) proposed an on-line version of the SHT procedure to control exposure rate at the item and test levels. Although the exposure parameters are still required for this on-line procedure to control item exposure rates and test overlap rate, the exposure parameters are updated sequentially on the fly, instead of through iterative simulations before operational test administration. This simulation study also showed that the on-line method performed better than the SHT procedure in controlling item exposure and test overlap at the early stage of the simulation (Chen, Lei, & Liao, 2008).

To control item exposure and test overlap in CATs more efficiently, Chen and Lei (2009) developed an on-line exposure control procedure that does not require any exposure parameters. Instead, this on-line method uses an item's usage status earlier in the testing to evaluate if this item is administered to the next examinee given that it is selected. Noting that the procedures previously developed to control item exposure and test overlap simultaneously were designed only for the context of item overlap between pairs of examinees, Chen and Lei (2009) extended Chen, et al.'s (2003) study by deriving a mathematical relationship between item exposure rates and test overlap rate for a group of examinees greater than two. Three forms of test overlap were defined in Chen and Lei's (2009) study: item sharing, unordered item pooling, and ordered item pooling. Item sharing was defined as the number of items commonly shared by all test takers in a group, unordered item pooling was the number of common items between an examinee and any possible α examinees in a group, and ordered item pooling referred to the number of overlapping items between a test taker and any α examinees who took the test earlier. Moreover, Chen and Lei (2009) suggested an exposure control procedure to control test overlap-ordered item pooling on the fly based on the notion that test overlap rate—ordered item pooling for the first t examinees is a function of test overlap rate—ordered item pooling for the previous (t-1) examinees.

In continuous testing such as CAT, controlling test overlap-ordered item pooling may be more practically useful than controlling the unordered item pooling (Chen & Lei, 2009) because continuous testing is ordered in the sense that tests are administered sequentially. However, this pro-

posed procedure has not been applied in any research or operational setting. A thorough investigation may be required to examine the effects of the proposed procedure on controlling test overlap rate. In automated test assembly with a heuristic such as the weighted deviations model (WDM) heuristic, test forms are assembled sequentially, which is similar to tests administered continuously. The exposure control procedure to manage test overlap-ordered item pooling on the fly appears to meet the needs of controlling test overlap rate for tests assembled sequentially. Additionally, estimating the exposure parameters for all items in the pool may not be necessary because not only is the number of assembled alternate forms much smaller than the number of examinees taking CATs, but also only items with particular characteristics would be selected for assembling alternate test forms. Therefore, controlling test overlap-ordered item pooling would be an appropriate method for automated test assembly and was applied in the current research. More detailed explanations of the ordered-item-pooling control are presented in the section to follow.

Purpose of the Study

Based on the rationale described in the previous paragraph, the purpose of this study is to comprehensively investigate the effectiveness of ordered-item-pooling control in automated test assembly with the weighted deviations model (WDM). This exposure control procedure was applied to minimize test overlap rate between pairs of test forms while producing tests that still meet content and psychometric constraints. To develop a better understanding of how well the ordered-item-pooling control method functions in automated assembly of multiple forms, this study evaluates its performance under different conditions of test length and test-content specifications. The performance under various conditions was compared to those from the corresponding baseline ATA conditions, where test overlap controls were not considered. The evaluation criteria include (i) the conformity to the test-assembly constraints, (ii) test parallelism in terms of the resultant psychometric properties, (iii) average test overlap rate, and (iv) distribution of item exposure rate.

Ordered-item-pooling control was implemented outside of the ATA process rather than including it as another constraint to avoid compromising the conformity to the test-assembly constraints. That is, this exposure control method was used to control the administration (or utilization) rate of an item after the item was selected based on the test-assembly algorithm.

Benchmark for Minimizing Test Overlap Rate

The target maximum test overlap rate is required in the ordered-item-pooling control procedure, and the specification of the target rate should be optimal in the sense that the target rate results in minimum test overlap while producing tests meeting test-assembly constraints. Chang and Zhang (2002) propose a theoretical lower bound of test overlap rate under the assumption of completely randomized item selection. The theoretical lower bound could be used as a criterion to evaluate the degree of test overlap. Chen, Ankenmann, & Spray (2003) further suggested that $r = (k \div n)$ could serve as the target rate to minimize test overlap, where k is the test length and n is the pool size. However, selecting items according to psychometric and non-psychometric constraints rather than drawing them at random, $r = (k \div n)$ may not be realistic (Chen, et al., 2003). An optimal benchmark or target for test overlap rate would be a value greater than $(k \div n)$. However, the value of $r = (k \div n)$ could be used as a starting point for constraining test overlap rate in the ordered-item-pooling control procedure. This is a starting value in the sense that it will be replaced in the process of specifying or searching for an optimal maximum average test overlap rate, given that an optimal target should meet the standard of fulfilling the requirement of the test-overlap rate of less than $r = (k \div n)$ while satisfying test-assembly specifications. The process of specifying optimal maximum average test overlap rate is described further in the methods section.

Lin (2008) suggested that an expected baseline test-overlap rate, E*(BTOR)* (Chen, Ankenmann, & Spray, 2003) could be an alternative target for constraining test overlap rate that would be less likely to compromise test quality because this index ensures that content specifications will be met. This suggestion makes sense because the constraint E*(BTOR)* is less strict than the value of $r = (k \div n)$, and content requirements must be satisfied to ensure content validity in most alternate-form assembly problems. However, similar to the value of $r = (k \div n)$, achieving the test-overlap rate of E*(BTOR)* while meeting test-assembly specifications may not be possible under strict test assembly conditions (e.g., small pool sizes and too many constraints), and therefore another target that is slightly greater than E*(BTOR)* would be more realistic. However, achieving the minimum test-overlap rate while meeting content specifications and ensuring content validity could be an appropriate criterion to examine the acceptability of the resultant test overlap rate in this study. Therefore, comparisons of *ATOR* and E*(BTOR)* were conducted to evaluate the effectiveness of the ordered-item-pooling control procedure in ATA. A large difference between *ATOR* and E*(BTOR)* would signal a possible problem in test security and indicate the necessity of a more stringent exposure control procedure.

Accordingly, it is helpful to introduce E*(BTOR)* in detail. Given that there are J content categories, j = 1, 2, ..., J, in an item pool, the item pool of size n is stratified into n_1, n_2, ..., n_J mutually exclusive partitions. In addition to the psychometric constraints, the test-assembly specifications require that m_1, m_2, ..., m_J items from each of these content categories appear on each assembled form and each form needs to be m items in length. The expected value of the baseline overlap rate or E*(BTOR)* for one content area will be shown first, and E*(BTOR)* for the entire test of length, m, is the sum of these expected values. Given that the random variable, Y, is the number of identical items between any two tests, $Y \div m$ is the observed overlap rate for those tests. Possible values for the random variable, Y could be 0, 1, ..., m. When Y = 0, there are no shared items between any two forms, whereas, when $Y = m$ the test forms are identical.

When only content constraints are imposed in ATA, m_j items will be drawn at random from each content area j. For each content area j, there are

$$\binom{C_j}{m_j}$$

possible combinations of the m_j items selected from the C_j items in that content area. Within each content area j, repeated draws will produce X_j common items shared by pairs of test forms, and X_j is distributed as a hyper-geometric random variable. Therefore,

$$prob(X_j = x) = \frac{\binom{m_j}{x}\binom{C_j - m_j}{m_j - x}}{\binom{C_j}{m_j}}$$

is defined as the probability that X_j, the number of common items in content area j, equals x, where x = 0, 1, ..., m_j. The expected value of X_j (Ross, 1976) for the j^{th} content area is given as follows:

$$E(X_j) = \sum_{x=0}^{m_j} \frac{\binom{m_j}{x}\binom{C_j - m_j}{m_j - x}}{\binom{C_j}{m_j}} x = \frac{m_j^2}{C_j}$$

The expected value of Y over all J content areas can be expressed as:

$$E(Y) = \sum_{j=1}^{J} E(X_j) = \sum_{j=1}^{J} \frac{m_j^2}{C_j}$$

and thereby the expected value of the *baseline test overlap rate* can be expressed as:

$$E\frac{Y}{m} = \frac{1}{m} \times E(Y) = \frac{1}{m} \sum_{j=1}^{J} \frac{m_j^2}{C_j}$$

Controlling Test Overlap—Ordered Item Pooling

For continuous testing, Chen and Lei (2009) express average test overlap rate—ordered item pooling for the t tests administered or assembled sequentially (i.e., $\overline{\Omega}_{\alpha,\, t}$) as a function of average test overlap rate—order item pooling for the previous (t-1) tests (i.e., $\overline{\Omega}_{\alpha,\, (t\text{-}1)}$). The expression is given in the following equation (please see Chen & Lei (2009) for the detailed derivation):

Equation 1:

$$\overline{\Omega}_{\alpha,\, t} = \left(\frac{t - \alpha - 1}{t}\right)(\overline{\Omega}_{\alpha,\, (t\text{-}1)}) + \frac{\alpha + 1}{t} - \frac{\sum_{i=1}^{n} [m_{it} - m_{i(t\text{-}1)}](\binom{t}{\alpha} - m_{it})}{m\,(\binom{t}{\alpha+1})}$$

where t is the number of tests administered or assembled, m_{it} is the number of times item i appears over the t tests, $m_{i(t\text{-}1)}$ is the number of times item i appears over the previous (t-1) tests, m is the test length and n is the number of items in the pool. If item i is included in the t^{th} test, the number of times item i appears across the t tests is greater than the number of times item i appears over the previous (t-1) tests by one and thereby $m_{it} - m_{(it\text{-}1)} = 1$. On the other hand, $m_{it} - m_{(it\text{-}1)} = 0$ when item i is not included in the t^{th} test. Therefore,

$$\sum_{i=1}^{n} [m_{it} - m_{i(t\text{-}1)}](\binom{t}{\alpha} - m_{it})$$

is the sum of

$$[m_{it} - m_{i(t\text{-}1)}](\binom{t}{\alpha} - m_{it})$$

over the m items in a test although the summation

$$\sum_{i=1}^{n}$$

is defined for all n items in the pool. This equation will yield the average percentage of common items between a particular test and a group of α tests that have been previously administered or assembled. If $\alpha = 1$, $\overline{\Omega}_{\alpha, t}$ represents the average test overlap rate between pairs of tests.

Example

Average test overlap rate—ordered item pooling can be computed based on the rate computed for previous tests using Equation 1 (previous page). Given that $\alpha = 1$, $m = 5$, and the number of test to be assembled (t) equals 5, to compute the average test overlap rate between pairs of tests for the five assembled tests ($\overline{\Omega}_{1, 5}$), the average test overlap rate between pairs of tests for the previous four assembled tests ($\overline{\Omega}_{1, 4}$) needs to be obtained. Similarly, $\overline{\Omega}_{1, 3}$ needs to be obtained to compute $\overline{\Omega}_{1, 4}$, and $\overline{\Omega}_{1, 2}$ needs to be obtained to compute $\overline{\Omega}_{1, 3}$. Table 1 shows the items selected in the tests, which are assembled sequentially. Along with the data in Table 1, Table 2 lists the number of times each item is selected for the first t assembled tests (m_{it}), $t = 1, 2, \ldots 5$. Take Test 3 ($t = 3$) for example, items 3, 6, 9, 2, and 4 are selected into the test (Table 1, next page). At this point, item 1 has not been used, item 2 is selected for the first time, item 9 has been selected three times, items 3, 4, 6, and 7 have been selected twice, and items 5, 8, and 10 have been selected once (Table 2, next page). Substituting the data from Table 2 into Equation 1:

$\overline{\Omega}_{1, 2} = {}^{2}/_{5}$, $\overline{\Omega}_{1, 3} = {}^{7}/_{15}$, $\overline{\Omega}_{1, 4} = {}^{8}/_{15}$, and $\overline{\Omega}_{1, 5} = {}^{14}/_{25}$.

Detailed computation is provided as follows:

$$\overline{\Omega}_{1, 2} = \frac{2}{5}$$

$$\overline{\Omega}_{1, 3} = \left(\frac{3-1-1}{3}\right)\left(\frac{2}{5}\right) + \frac{1+1}{3} - \frac{\binom{3-1}{1}+\binom{3-2}{1}+\binom{3-2}{1}+\binom{3-2}{1}+\binom{3-3}{1}}{5\binom{3}{1+1}} = \left(\frac{7}{15}\right)$$

$$\overline{\Omega}_{1, 4} = \left(\frac{4-1-1}{4}\right)\left(\frac{7}{15}\right) + \frac{1+1}{4} - \frac{\binom{4-3}{1}+\binom{4-3}{1}+\binom{4-2}{1}+\binom{4-4}{1}+\binom{4-2}{1}}{5\binom{4}{1+1}} = \left(\frac{8}{15}\right)$$

$$\overline{\Omega}_{1, 5} = \left(\frac{5-1-1}{5}\right)\left(\frac{8}{15}\right) + \frac{1+1}{5} - \frac{\binom{5-4}{1}+\binom{5-2}{1}+\binom{5-3}{1}+\binom{5-5}{1}+\binom{5-3}{1}}{5\binom{5}{1+1}} = \left(\frac{14}{25}\right)$$

Table 1: Items Included in Each Assembled Test

Test number *(t)*	Item *(i)* selected into test				
1	3	4	7	6	9
2	9	5	7	8	10
3	3	6	9	2	4
4	9	8	10	4	3
5	7	4	5	10	9

Table 2: Item Usage Corresponding to Items Selected in Table 1 for the First *t* Tests

	Item *(i)* selected into test									
Test number *(t)*	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	0	0	1	1	0	1	1	0	1	0
2	0	0	1	1	1	1	2	1	2	1
3	0	1	2	2	1	2	2	1	3	1
4	0	1	3	3	1	2	2	2	4	2
5	0	1	3	4	2	2	3	2	5	3

Chen and Lei (2009) pointed out a critical element in controlling $\overline{\Omega}_{\alpha,\, t}$, which is

$$\lambda = \sum_{i=1}^{n} [m_{it} - m_{i(t-1)}](\overset{t}{\alpha} - m_{it})$$

as α, t, m, and $\overline{\Omega}$ are known after (t-1) tests have been administered or assembled. For the exposure-control procedure based on Equation 1 (page 11), a maximum expected test overlap rate, $\overline{\Omega}_{\max}$, is first specified. To constrain $\overline{\Omega}_{\alpha,\, t}$ at a rate no greater than $\overline{\Omega}_{\max}$, λ and $\overline{\Omega}_{\max}$ would have the following relationship.

Equation 2:

$$\overline{\Omega}_{\alpha,\, t} = \left(\frac{t - \alpha - 1}{t}\right)(\overline{\Omega}_{\alpha,\, (t-1)}) + \frac{\alpha + 1}{t} - \frac{\sum_{i=1}^{n} [m_{it} - m_{i(t-1)}](\overset{t}{\alpha} - m_{it})}{m\,(\overset{t}{\alpha+1})} \leq \overline{\Omega}_{\max}$$

Equation 3: $$\lambda = \sum_{i=1}^{n} [m_{it} - m_{i(t-1)}]\binom{t-m_{it}}{\alpha} \geq m \binom{t}{\alpha+1} \left[\frac{t-\alpha-1}{t} \times \overline{\Omega}_{\alpha,\,(t-1)} + \frac{\alpha+1}{t} - \overline{\Omega}_{\max} \right] = D$$

As described previously, λ is the sum of

$$[m_{it} - m_{i(t-1)}]\binom{t-m_{it}}{\alpha}$$

over the m items in a test although equation 3 defines λ as the sum of

$$[m_{it} - m_{i(t-1)}]\binom{t-m_{it}}{\alpha}$$

over all n items in a pool. Based on Equation 3, to make λ no less than D, each item selected for inclusion should have a contribution of at least D/m added to λ, given that the contribution of each item refers to

$$[m_{it} - m_{i(t-1)}]\binom{t-m_{it}}{\alpha}$$

Note that the term $\overline{\Omega}_{\alpha,\,(t-1)}$ in D is computed based on the procedure described in the previous example section. As a result, D/m serves as an item inclusion criterion to govern the inclusion or administration of items. Specifically, to control $\overline{\Omega}_{\alpha,\,t}$ at a rate less than $\overline{\Omega}_{\max}$, an item should have contribution,

$$[m_{it} - m_{i(t-1)}]\binom{t-m_{it}}{\alpha}$$

no less than D/m to be included in the t^{th} test. This criterion for item inclusion to control test overlap rate seems reasonable in ATA given that a greater $(t - m_{it})$ signals a smaller number of times that item i is included in or appears over the t tests, and thereby items with greater

$$[m_{it} - m_{i(t-1)}]\binom{t-m_{it}}{\alpha}$$

would tend to be included in a test currently assembled to make item-usage distribution more even. Based on the previous example, if $\Omega_{\max}$ is specified as 0.2 and the 4th test is to be assembled, Equation 3 becomes

$$\lambda = \sum_{i=1}^{10} [m_{i4} - m_{i(4-1)}]\binom{4-m_{it}}{1} \geq 5 \binom{4}{1+1} \left[\frac{4-1-1}{4} \times \frac{7}{15} + \frac{1+1}{4} - 0.2 \right] = D$$

or $\lambda \leq D = 16$. To control $\overline{\Omega}_{\alpha,\,t}$ (i.e., $\overline{\Omega}_{1,\,4}$) at a rate no greater than $\overline{\Omega}_{\max}$, an item with contribution,

$$[m_{i4} - m_{i(4-1)}]\binom{4-m_{it}}{1}$$

no less than $16/5$ is included in the test. Similarly, when the 5th test is to be assembled, Equation 3 becomes

$$\lambda = \sum_{i=1}^{10} [m_{i5} - m_{i(5-1)}]\binom{5-m_{it}}{1} \geq 5 \binom{5}{1+1} \left[\frac{5-1-1}{5} \times \frac{8}{15} + \frac{1+1}{5} - 0.2 \right] = D$$

or $\lambda \geq D = 26$. An item with contribution,

$$[m_{i5} - m_{i(5\text{-}1)}](\tfrac{5}{1}-m_{it})$$

not less than $^{26}/_{5}$ is included in the test to control $\overline{\Omega}_{1,\,5}$ not greater than 0.2.

Methods

The following section introduces the design of the study, characteristics of the item pool, properties of the reference target on which to base the test-assembly constraints, procedure of implementing the ordered-item-pooling control in ATA with the WDM heuristic, specification of maximum average test overlap rate, and criteria used to evaluate the effectiveness of the ordered-item-pooling control.

Design

The ordered-item-pooling control procedure was compared to the baseline condition—no exposure control, to investigate the effectiveness of this exposure control method. Variables examined that might influence this comparison included (1) ratio of test length to the size of item pool, and (2) test-content specification. Therefore, this comparison was conducted under the experimental conditions from the combination of the two variables. To investigate the effectiveness of the ordered-item-pooling control process for short, moderate, and long tests, three test lengths of 60, 90, and 120 items respectively were studied. Three test lengths represented three ratios of test length to the size of item pool, which were 0.10 (60 ÷ 600), 0.15 (90 ÷ 600), and 0.20 (120 ÷ 600). In this study, because only one item pool was used, test length and the ratio of test length to the pool size were used interchangeably.

Additionally, two content specifications were studied. One content specification mirrored the content distribution of the item pool and is called balanced content specification. The other did not match the content distribution and is called unbalanced content specification. Accordingly, the effectiveness of the ordered-item-pooling control process was evaluated under 6 experimental conditions (i.e., 3 ratios of test length to the size of item pool × 2 content specifications). Ten test forms were assembled under each experimental condition using the WDM heuristic—a flexible and straightforward method for automated test assembly.

Item Pool

The item pool used in this study contained 600 calibrated Mathematics items, and covered four content areas: 152 content-A items, 127 content-B items, 147 content-C items, and 174 content-D items (Table 3). The 600 mathematics items were calibrated using the 3-parameter logistic model (3-PLM) in the computer program, Bilog-MG (Zimowski, Muraki, Mislevy, & Bock, 2003) with the default prior $\theta \sim N(0,1)$. The average item parameters, $\overline{a}$, $\overline{b}$, and $\overline{c}$, in the pool were 0.965, 0.127, and 0.178, respectively. Additionally, the average item p-values and point-biserial correlations in the pool were 0.559, and 0.395, respectively. The item pool contained predominantly medium to slightly difficult items as the test information for the Mathematics item pool peaked at $\theta = 1$.

Table 3: Content Distribution of the Item Pool and Test-content Specification for Short, Mid-length, and Long Tests (ratio = 0.1, 0.15, and 0.2, respectively) Under Balanced and Unbalanced Content Conditions

		Ratio = 0.1 = 60/600		Ratio = 0.15 = 90/600		Ratio = 0.2 = 120/600	
Content Area	Item Pool	Balanced	Unbalanced	Balanced	Unbalanced	Balanced	Unbalanced
A	152 (25%)	15 (25%)	8 (13%)	23 (25%)	12 (13%)	30 (25%)	16 (13%)
B	127 (21%)	13 (21%)	4 (7%)	19 (21%)	6 (7%)	25 (21%)	8 (7%)
C	147 (25%)	15 (25%)	20 (33%)	22 (25%)	30 (33%)	30 (25%)	40 (33%)
D	174 (29%)	17 (29%)	28 (47%)	26 (29%)	42 (47%)	35 (29%)	56 (47%)

Note: ratio = ratio of test length to item-pool size.

ATA Constraints

To investigate how the ordered-item-pooling control process performed under various test lengths and content specifications, three test lengths (short, mid-length, and long) and two content specifications (balanced and unbalanced) were studied. One content specification mirrored the content distribution of the item pool, but the other did not. The percentages of content areas for the two sets of test-content specifications are presented in Table 3. The combinations of test length and content specification yielded six reference targets (Table 4, next page). The first reference target represented a short target test with content distribution mirroring that of the pool (i.e., balanced content outline); the second reference target represented a short target test with content distribution not mirroring that of the pool (i.e., unbalanced content outline); the third

reference target represented a mid-length target test with content distribution mirroring that of the pool; the fourth reference target was a mid-length target test with content distribution not mirroring that of the pool; the fifth reference target was a long target test with content distribution mirroring that of the pool; the sixth reference target was a long target test with content distribution not mirroring that of the pool.

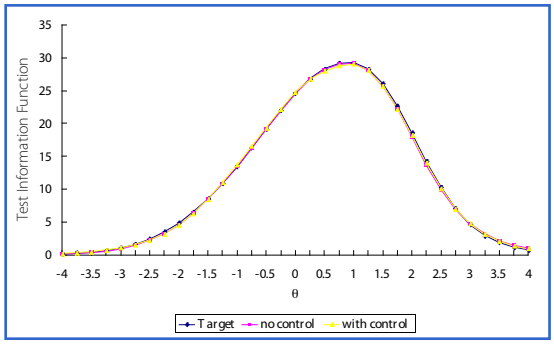
Table 4: Six Reference Targets

	Ratio = 0.1 = 60/600	Ratio = 0.15 = 90/600	Ratio = 0.2 = 120/600
Content distribution mirrors the pool	1st Reference	3rd Reference	5th Reference
Content distribution does *not* mirror the pool	2nd Reference	4th Reference	6th Reference

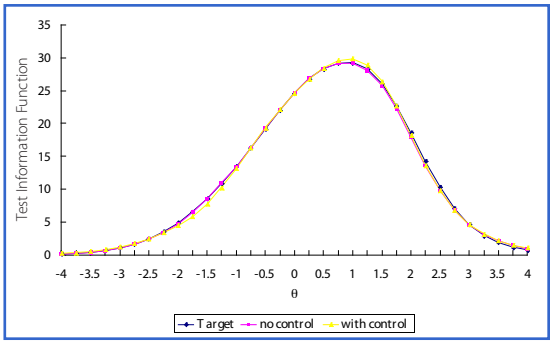
Note: ratio = ratio of test length to item-pool size.

These reference targets were used to define the content and psychometric targets required for all constructed forms of the test. Under a particular content specification, the percentage of content specifications was the same for all test lengths. In terms of psychometric properties, each of the ten tests assembled was constrained to have the test information function matching the target function from the reference target. Within each reference target, the target test information was specified at 33 θ points, ranging from –4.00 to +4.00 in increments of 0.25. Under a given test length, the target test information function was the same for the balanced and unbalanced content specifications. That is, the psychometric constraints varied with test length. The target test information functions from the six reference forms were plotted in Figures 1–3 (below and next page).

Figure 1: Test Information Functions for the 60-Item Reference Test and Poorest Matching Test Forms With and Without Test-overlap Control Under *Balanced* (I) and *Unbalanced* (II) Content Conditions

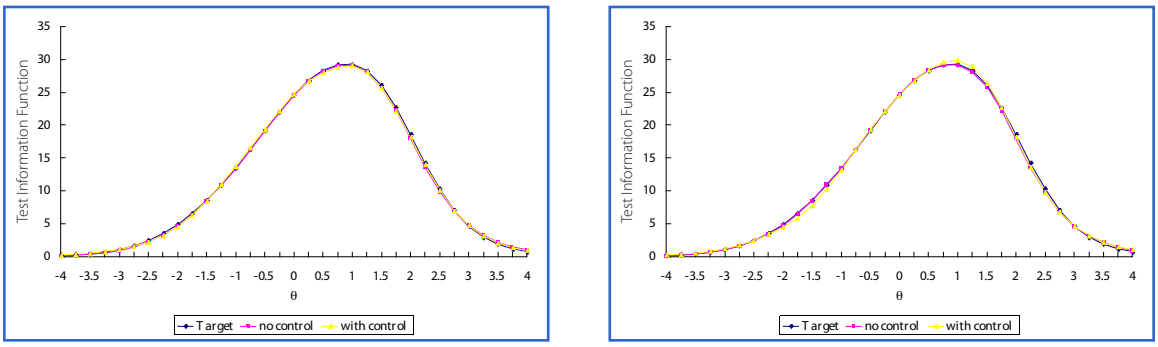


I: Balanced



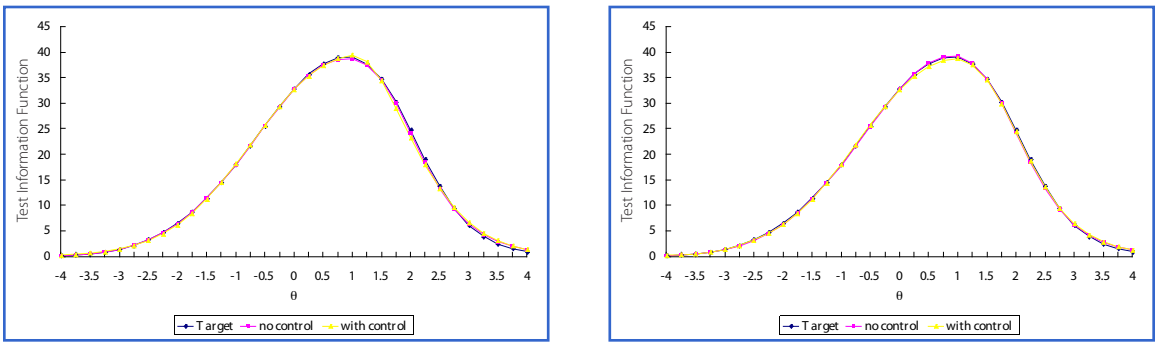
II: Unbalanced

Figure 2: Test Information Functions for the 90-Item Reference Test and Poorest Matching Test Forms With and Without Test-overlap Control Under *Balanced* (I) and *Unbalanced* (II) Content Conditions



I: Balanced II: Unbalanced

Figure 3: Test Information Functions for the 120-Item Reference Test and Poorest Matching Test Forms With and Without Test-overlap Control Under *Balanced* (I) and *Unbalanced* (II) Content Conditions



I: Balanced II: Unbalanced

Overall, the ATA task was to assemble ten tests with the WDM heuristic that followed the content and psychometric constraints specified from each reference form with all constraints weighted equally. Analyses were done following the content and psychometric constraints for six separate reference targets. In practice, the constraints may consist of upper and lower boundaries around the target values, so that there is some degree of flexibility in meeting each constraint. For this study, 37 constraints were defined to assemble a test form under each reference target, including 4 content and 33 psychometric constraints.

The ATA constraints specified in this study reflect a typical scenario of assembling equivalent forms. Within the framework of IRT and automated test assembly, alternate test forms are typically produced based

on Samejima's definition of weakly parallel forms in which the forms are matched to a target test information function (TTIF) (e.g., Armstrong, Jones, Li, & Wu, 1996; Luecht & Hirsch, 1992; Luecht, 1998; Swanson & Stocking, 1993; van der Linden & Boekkooi-Timminga, 1989; van der Linden & Adema, 1997). This is a method reasonably simple to implement because item information functions are additive and easy to manipulate (van der Linden, personal communication).

Specification of Maximum Average Test Overlap Rate

For ATA, conformity to the test-assembly constraints is the major concern. The optimal exposure control procedure for ATA would be the method that generates a minimum average test overlap rate *(ATOR)* while meeting all assembly requirements or constraints, where "minimum" must be specified. In the ordered-item-pooling control process, the maximum expected test overlap rate, $\overline{\Omega}_{max}$, should be first specified so that the resultant average test overlap rate would not exceed $\overline{\Omega}_{max}$. Accordingly, choosing a $\overline{\Omega}_{max}$ such that the resultant average test overlap rate is a minimum and all assembly constraints have been satisfied is an important task in this exposure control procedure for ATA.

Under each condition of the study, the value of $\overline{\Omega}_{max}$ for ten test forms was set to $(k \div n)$, which is a theoretical lower bound of test overlap rate under the assumption of completely randomized item selection, at the beginning of a set of iterations, where ten test forms were assembled at each iteration. The value of $\overline{\Omega}_{max}$ was increased on successive computer runs until a value of $\overline{\Omega}_{max}$ produced ten forms that met all test-assembly constraints and yielded a minimum value for *ATOR*. The $\overline{\Omega}_{max}$ value obtained at the last iteration was the optimal $\overline{\Omega}_{max}$ and was the maximum expected test overlap rate to be specified in the ordered-item-pooling control process. Only a few iterations are necessary to obtain the optimal $\overline{\Omega}_{max}$.

Procedure of Assembling Alternate Test Forms Using the WDM Heuristic with Ordered-Item-Pooling Control

The test assembly algorithm used for this study was the weighted deviations model (WDM) heuristic developed by Swanson and Stocking (1993). The WDM heuristic procedure designed for automated test assembly can be categorized as a greedy heuristic algorithm. The goal of the greedy heuristic is the pursuit of the greatest improvement at each iteration toward an optimal solution. Accordingly, with the WDM heuristic, items are selected sequentially so that those chosen first provide the best improvement in conforming to all the constraints simultaneously.

The test overlap control method applied in this study was the ordered-item-pooling control procedure. To ensure that the average test overlap

rate was less than $\overline{\Omega}_{max}$, an item was included in a test if it was selected according to its ability to satisfy all of the constraints imposed in WDM and had contribution,

$$[m_{it} - m_{i(t-1)}](t_{\alpha}^{-}m_{it})$$

of no less than the value of $^{D}/_{m}$. The procedure of assembling alternate test forms using the WDM heuristic with the ordered-item-pooling control is listed in the following section.

1. Determine the content constraints (i.e., test-content specification), psychometric constraints, the test length, n, the number of test forms to be drawn, and the weights for each of the constraints.
2. Specify the maximum average test overlap rate, $\overline{\Omega}_{max}$ (described in the previous section).
3. Randomly select the first item. (Note: If multiple forms are to be drawn, the first item entering the test should be randomly selected from the item pool to avoid the same test form being assembled repeatedly (Parshall, Spray, Kalohn, & Davey, 2001). If the same item is selected to enter the test first, the same test form will be constructed repeatedly because the item-selection process based on the WDM heuristic is affected by past as well as future item selections.)
4. Evaluate items sequentially in terms of their ability to fulfill all the constraints with WDM.
5. Select the item with the greatest ability to fulfill the constraints with WDM.
6. Compare this selected item's contribution,

   $[m_{it} - m_{i(t-1)}](t_{\alpha}^{-}m_{it})$

   to the value of $^{D}/_{m}$.
7. Evaluate if its

   $[m_{it} - m_{i(t-1)}](t_{\alpha}^{-}m_{it})$

   is not less than the value of $^{D}/_{m}$. If yes, include this item on the test form. Otherwise, set aside this item and select the next best item and evaluate if the corresponding

   $[m_{it} - m_{i(t-1)}](t_{\alpha}^{-}m_{it})$

   is not less than the value of $^{D}/_{m}$.
8. Repeat Step (7) until the qualified item appears and include it on the test form being assembled.

9. Repeat steps 4–8 until all *m* items (i.e., test length) are selected.
10. Repeat steps 3–9 until the specified number of alternate test forms is reached.

Evaluation Criteria

In ATA, an exposure control method would need to demonstrate its effectiveness by achieving an acceptable average test overlap rate across multiple forms without compromising the conformity to the test-assembly constraints. Additionally, test equivalence of the assembled forms in the resultant psychometric properties (e.g. test difficulty) would need to be evaluated because the ATA task in the study involved equivalent-form assembly. Therefore, comparing the results to the baseline condition where no exposure control was conducted, the research evaluated the performance of the ordered-item-pooling control procedure in terms of (i) the conformity to the content and psychometric constraints, (ii) test equity among the assembled forms in the resultant psychometric properties, (iii) average test overlap rate, and (iv) distribution of item exposure rate. Specifically, whether an average test overlap rate across multiple forms was acceptable was examined in terms of criteria (iii) and (iv).

Test Equity of the Assembled Forms

An assembled test is parallel to the reference test in content if it has the same content distribution as the reference test. A measure of this conformity is the percentage of content specifications met. Additionally, the assembled forms were constrained to have the TIF equivalent to a target TIF in the study, and thereby it is necessary to examine the extent to which the TIFs of the resultant test forms are similar to the target TIF.

To evaluate if the addition of the ordered-item-pooling control method into the ATA process would affect the equity of the assembled forms in the resultant psychometric properties, several indices conditional on the proficiency scales were examined among the assembled forms under each experimental condition. These conditional indices included the first central moment of $P(\theta)$ or test characteristic function (i.e., TCF), the square root of second central moment of $P(\theta)$, and the conditional error variance of observed test score X (i.e., CEV of X). $P(\theta)$ denotes the item characteristic function. These psychometric properties are resultant in the sense that they are the outcomes rather than test-assembly constraints in the ATA process.

The first central moment of $P(\theta)$ is critical in assessing the conditional difficulty of the assembled tests. It is also important to evaluate the simi-

larity among the assembled tests in the second central moment of $P(\theta)$ to ensure equivalence in the variability of item difficulty over a span of proficiency points. The conditional error variance of observed test scores is defined as the sum of the product of the conditional correct-response probability and the complement probability over all items in a test at selected proficiency points. Evaluating the congruence of the CEV of X among the assembled tests is important for ensuring the equivalence of conditional measurement error at selected proficiency points.

Average Test Overlap Rate *(ATOR)*

Test overlap rate is an important factor to consider in ensuring the security of test items. In this research, test overlap rate is defined as the percentage of items shared by a pair of assembled forms of a fixed length. The average test overlap rate between pairs of constructed test forms can be obtained by computing the percentage of test overlap for all possible pair-wise constructed forms, and then taking the average over all of these percentages.

In this study, *ATOR* was computed and compared to an expected baseline test overlap rate, E*(BTOR)*. E*(BTOR)* is the test-overlap rate when only the content constraints are imposed for automated test assembly.

Distribution of Item Utilization

In computerized adaptive testing (CAT), the exposure rate of an item is defined as the percentage of all CAT administrations in which the item is included. In automated test assembly, the exposure rate of an item refers to the utilization rate of an item, which is defined as the ratio of the number of times the item is selected into test forms over the total number of assembled test forms. Moreover, the number of times the item is selected into test forms could be rephrased as the number of forms in which the item appears or "item utilization" in short.

In this study, the distribution of item utilization was examined to evaluate if the ordered-item-pooling control procedure was effective in reducing item overexposure and enhancing pool utilization. Over-exposed items and un-utilized items would damage test security and the cost effectiveness of the item pool, respectively. The smaller the number of over-exposed items and un-utilized items or the more even the item-utilization distribution, the better the ordered-item-pooling control method performs. In this study, special emphasis was placed on the comparison of the no exposure control and the ordered-item-pooling control procedures in terms of the maximum item exposure rate and the number of unused items.

Results

Conformity to the Automated-Test-Assembly (ATA) Constraints

In this study, the value of $\overline{\Omega}_{max}$ was determined such that it would result when the average test overlap rate was a minimum and all assembly constraints have been satisfied. Therefore, to ensure valid interpretation of the results, it is important to first consider the extent to which the imposed constraints were fulfilled. Once the automated-test-assembly (ATA) process was completed, the constructed tests were evaluated by examining the extent to which they satisfied the constraints imposed under each condition. Figures 1–3 (pages 17 and 18) show the information plots for the target test and the poorest matching test form generated under test length = 60, 90, and 120, respectively. Each figure includes two graphs—the left graph (graph I) is plotted for the no exposure control condition while the right graph (graph II) is plotted for the exposure control condition. These figures show that the information plots for the generated and target tests were very similar throughout most regions of the proficiency scale with somewhat better conformity to constraints for the no overlap control condition than for the test overlap control condition. When differences in test information occurred between the target and generated tests, they were negligible in most cases. These deviations were not considered substantial enough to invalidate the subsequent results that will be discussed shortly. Content constraints were met for all reference tests. Therefore, all assembled test forms produced 100% content parallelism.

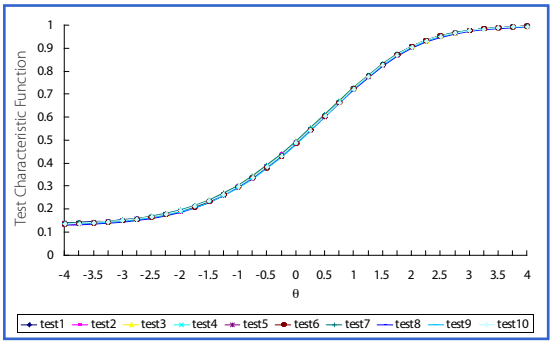
Test Equity in Terms of the Resultant Psychometric Properties

In this section, the comparability among alternate forms assembled within each experimental condition on the first and second central moments of $P(\theta)$, and the error variance of observed test scores is examined conditionally at selected proficiency points. The corresponding results are plotted in Figures 4–21. Similarly, each figure includes two graphs with the left graph (graph I) representing the no exposure control condition and the right graph (graph II) representing the exposure control condition. Each graph shows the curves for the 10 test forms created in each experimental condition, where smaller variation over the 10 curves indicates a higher degree of equivalence of test forms and vice versa.

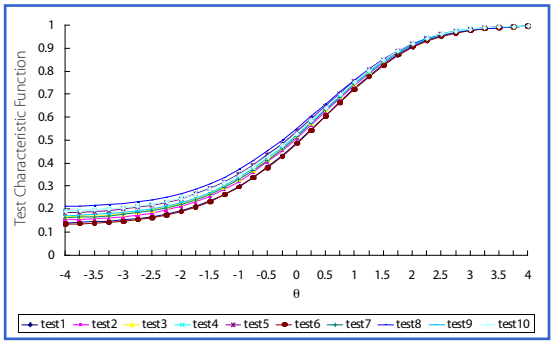
The test characteristic function (TCF) or first central moment of $P(\theta)$ is defined as the average of the conditional correct-response probabilities over all items in a test at selected θ values. Figures 4–9 (pages 24 to 26) show the test characteristic function (TCF) plot for the 10 test forms

created in each experimental condition. The graph at the left hand side of each Figure (i.e., graph I) plots the TCFs of the 10 test forms created under the no exposure control condition, whereas the graph at the right hand side (i.e., graph II) displays the TCFs under the exposure control condition. From these figures, the variations among the assembled tests occurred in the lower half regions of the proficiency scale (i.e., about $\theta \leq 0.00$) of the right graphs (i.e., graphs II), and the TCF variations in the left graphs (i.e., graphs I) tended to be smaller than those in the right graphs. A general impression conveyed from the plots is that the assembled tests without test overlap control were more similar than were the tests with test overlap control, given that smaller TCF variation indicates a higher degree of equivalence of test forms in TCF and vice versa. These results appeared to generalize across the test length (60 vs. 90 vs. 120) and content (balanced vs. unbalanced) conditions.

Figure 4: Test Characteristic Functions for *Balanced Without* (I) and *With* (II) Test Overlap Control for 60-Item Tests

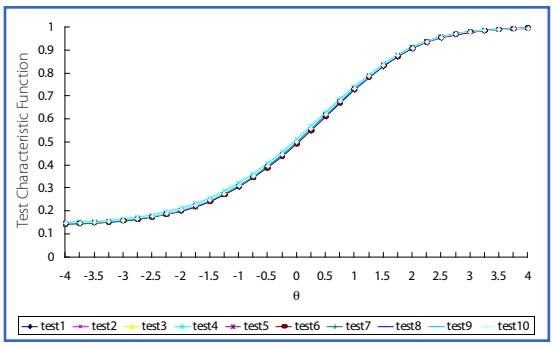


I: Without Test Overlap Control

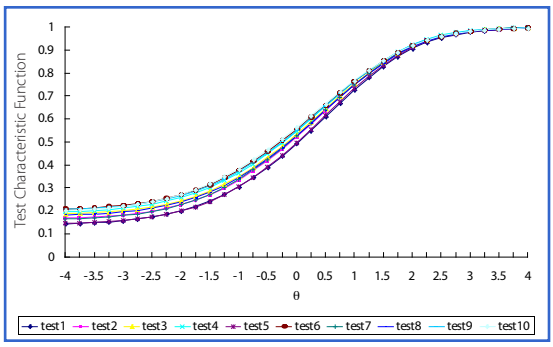


II: With Test Overlap Control

Figure 5: Test Characteristic Functions for *Balanced Without* (I) and *With* (II) Test Overlap Control for 90-Item Tests

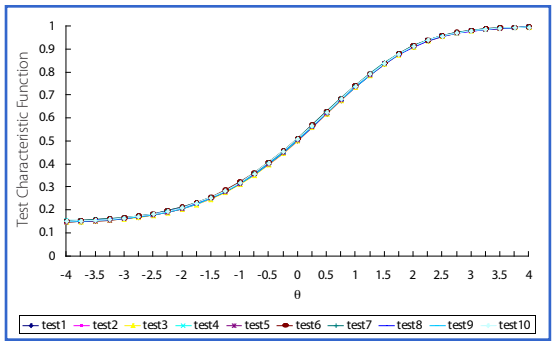


I: Without Test Overlap Control

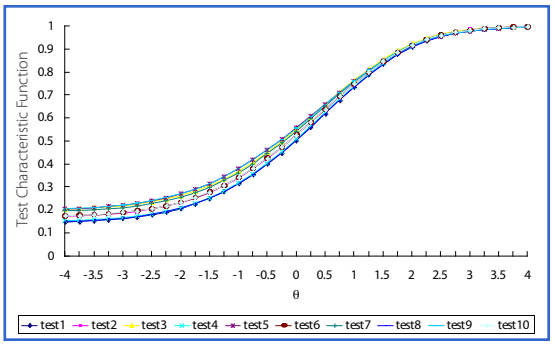


II: With Test Overlap Control

Figure 6: Test Characteristic Functions for *Balanced Without* (I) and *With* (II) Test Overlap Control for 120-Item Tests

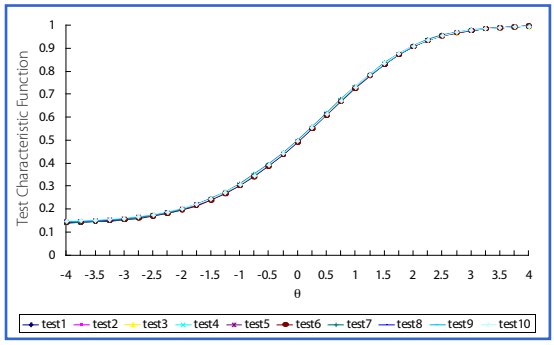


I: Without Test Overlap Control

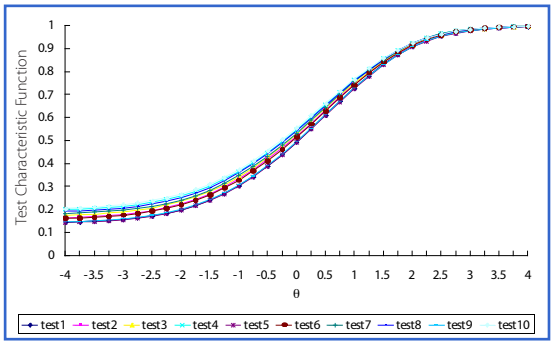


II: With Test Overlap Control

Figure 7: Test Characteristic Functions for *Unbalanced Without* (I) and *With* (II) Test Overlap Control for 60-Item Tests

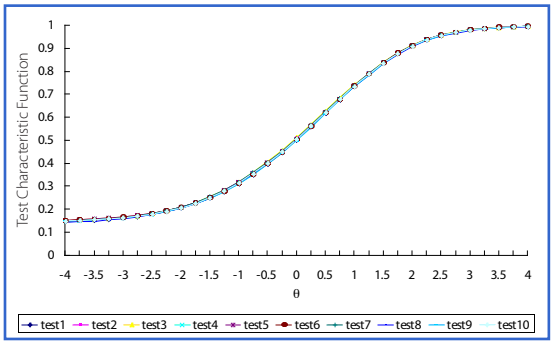


I: Without Test Overlap Control

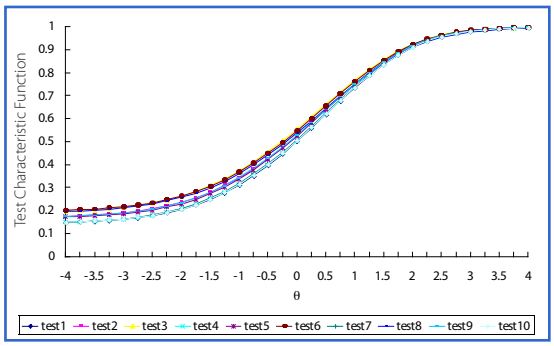


II: With Test Overlap Control

Figure 8: Test Characteristic Functions for *Unbalanced Without* (I) and *With* (II) Test Overlap Control for 90-Item Tests

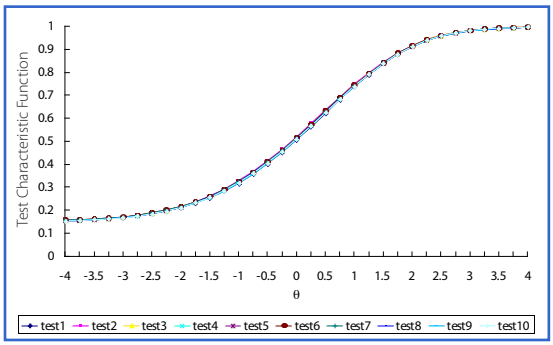


I: Without Test Overlap Control

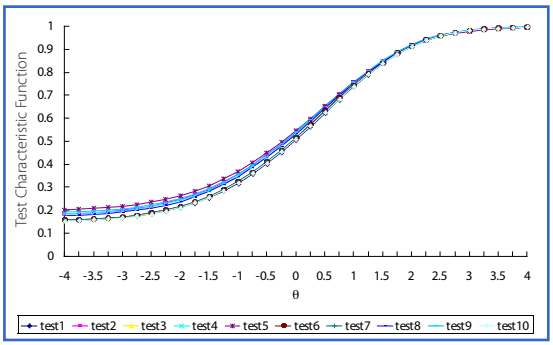


II: With Test Overlap Control

Figure 9: Test Characteristic Functions for *Unbalanced Without* (I) and *With* (II) Test Overlap Control for 120-Item Tests



I: Without Test Overlap Control

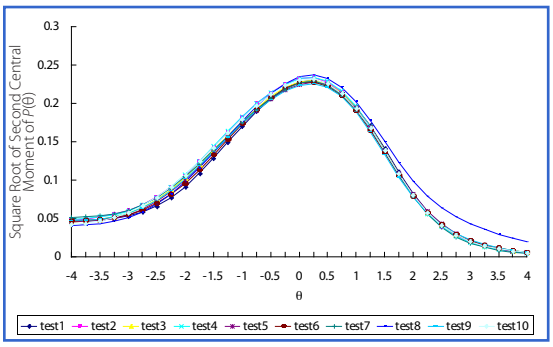


II: With Test Overlap Control

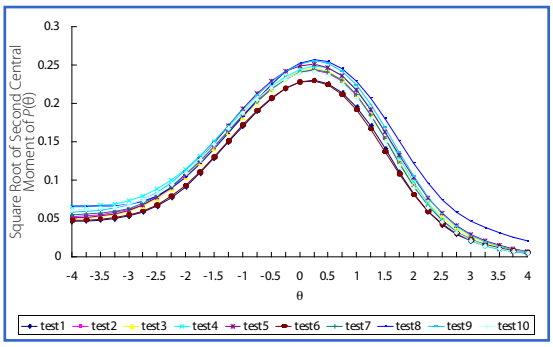
To provide a further evaluation of differences in TCFs among the assembled tests, the standard deviation (SD) of TCF across the 10 test forms was aggregated over all proficiency points in each experimental condition. This index provides an indicator of TCF variation aggregated over the entire proficiency scale. As shown in Table 5 (page 31), lower aggregated deviations tend to be associated with tests assembled without exposure control (range from 0.002 to 0.003), and higher deviations are associated with tests generated with exposure control (range from 0.012 to 0.017). These findings are consistent with those revealed in the previous TCF plots.

Figures 10–15 (pages 27 to 28) show plots of the square root of the second central moment of $P(\theta)$, and Figures 16–21 (pages 29 to 30) present plots of the conditional error variance of observed test score X (i.e., CEV of X) for the 10 test forms created in each experimental condition. Similarly, the left hand side of each Figure (i.e., graph I) displays the graph under the no exposure control condition, whereas the right hand side (i.e., graph II) presents the graph under the exposure control. Greater variation was observed for the right graphs. Therefore, the degree of test form equivalence in the second central moment of $P(\theta)$ and the CEV of X tended to be lower as the test overlap control was imposed. Once again, these results appeared to generalize across the test length (60 vs. 90 vs. 120) and content (balanced vs. unbalanced) conditions.

Figure 10: Square Root of Second Central Moment of *P*(θ) for 60-Item Tests Under *Balanced* Content Outline *Without* (I) and *With* (II) Test Overlap Control

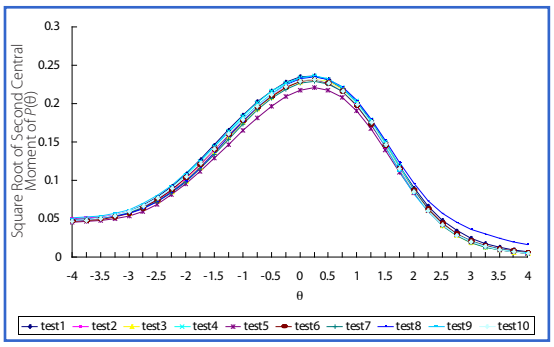


I: Without Test Overlap Control

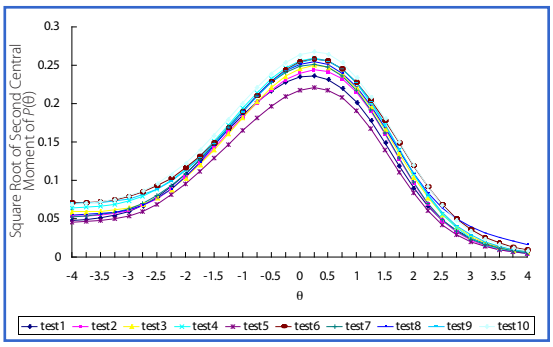


II: With Test Overlap Control

Figure 11: Square Root of Second Central Moment of *P*(θ) for 90-Item Tests Under *Balanced* Content Outline *Without* (I) and *With* (II) Test Overlap Control

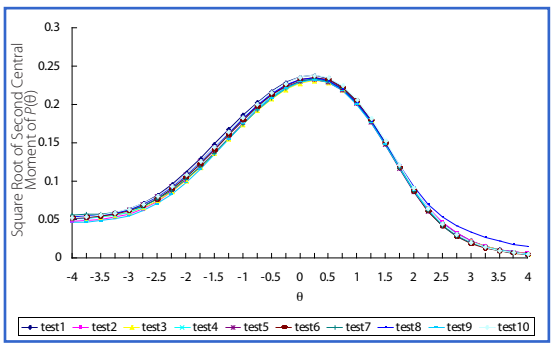


I: Without Test Overlap Control

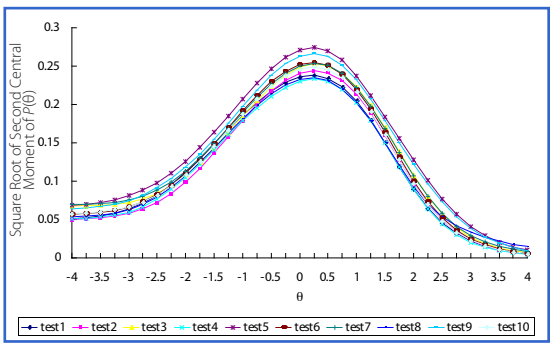


II: With Test Overlap Control

Figure 12: Square Root of Second Central Moment of *P*(θ) for 120-Item Tests Under *Balanced* Content Outline *Without* (I) and *With* (II) Test Overlap Control

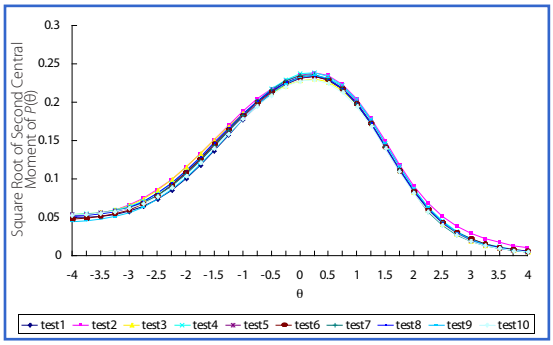


I: Without Test Overlap Control

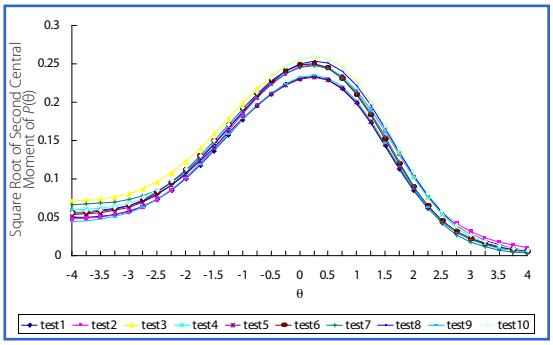


II: With Test Overlap Control

Figure 13: Square Root of Second Central Moment of *P*(θ) for 60-Item Tests Under *Unbalanced* Content Outline *Without* (I) and *With* (II) Test Overlap Control

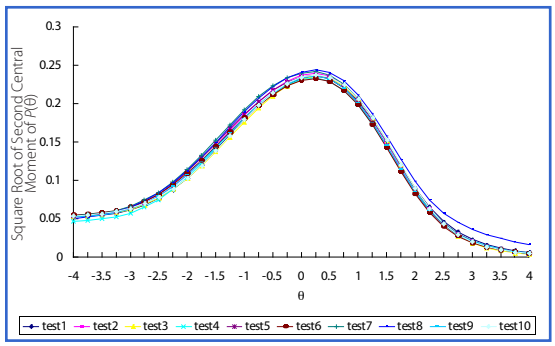


I: Without Test Overlap Control

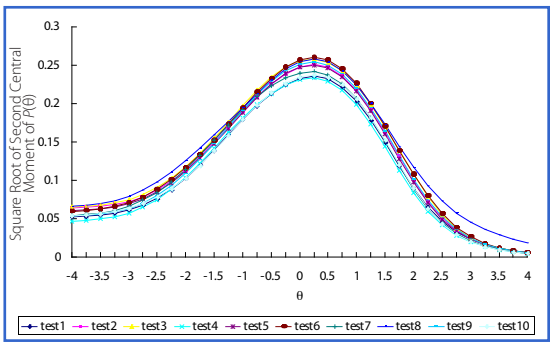


II: With Test Overlap Control

Figure 14: Square Root of Second Central Moment of *P*(θ) for 90-Item Tests Under *Unbalanced* Content Outline *Without* (I) and *With* (II) Test Overlap Control

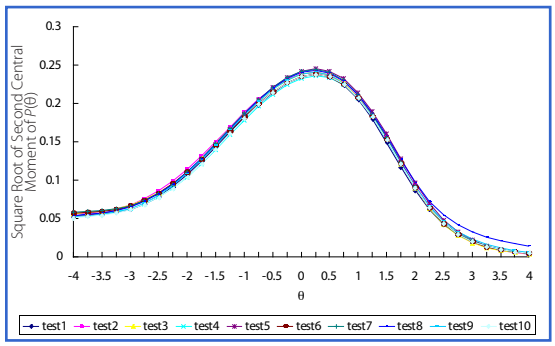


I: Without Test Overlap Control

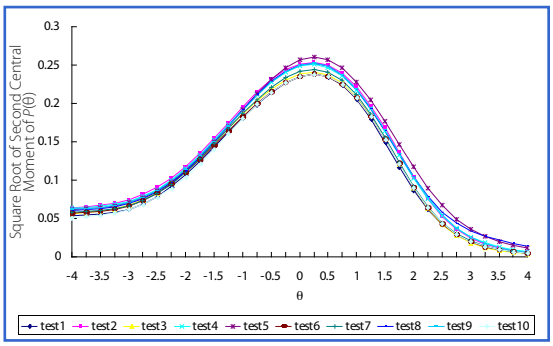


II: With Test Overlap Control

Figure 15: Square Root of Second Central Moment of *P*(θ) for 120-Item Tests Under *Unbalanced* Content Outline *Without* (I) and *With* (II) Test Overlap Control

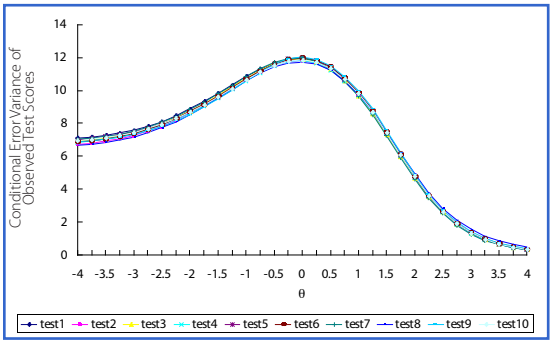


I: Without Test Overlap Control

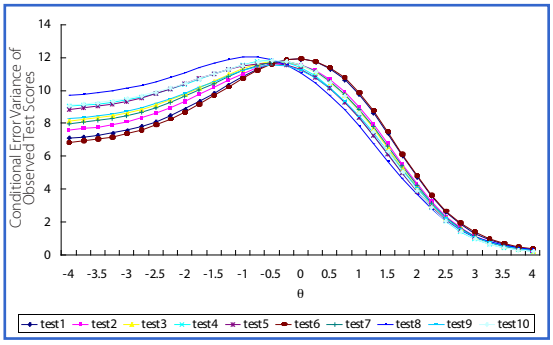


II: With Test Overlap Control

Figure 16: Conditional Error Variance of Observed Test Score (X) for 60-Item Tests Under *Balanced* Content Outline *Without* (I) and *With* (II) Test Overlap Control

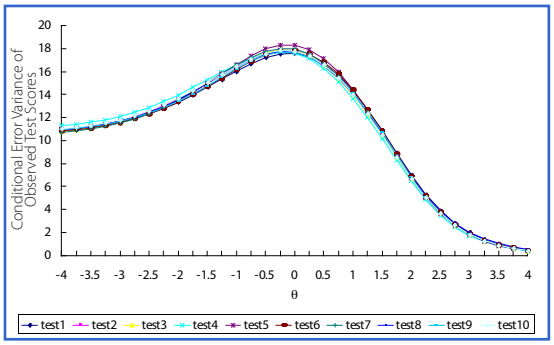


I: Without Test Overlap Control

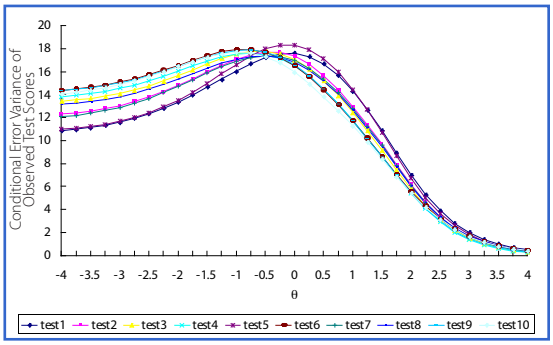


II: With Test Overlap Control

Figure 17: Conditional Error Variance of Observed Test Score (X) for 90-Item Tests Under *Balanced* Content Outline *Without* (I) and *With* (II) Test Overlap Control

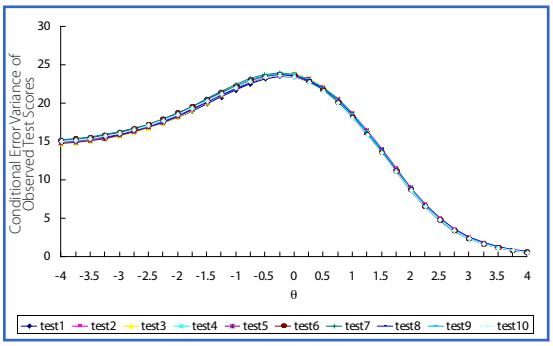


I: Without Test Overlap Control

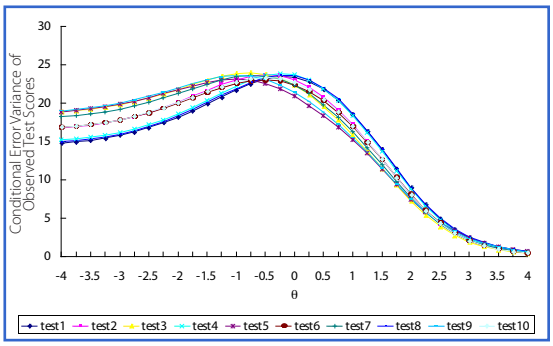


II: With Test Overlap Control

Figure 18: Conditional Error Variance of Observed Test Score (X) for 120-Item Tests Under *Balanced* Content Outline *Without* (I) and *With* (II) Test Overlap Control

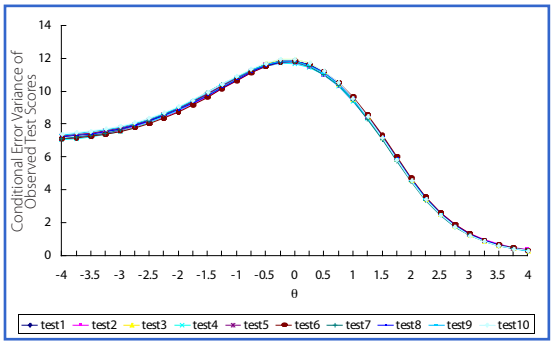


I: Without Test Overlap Control

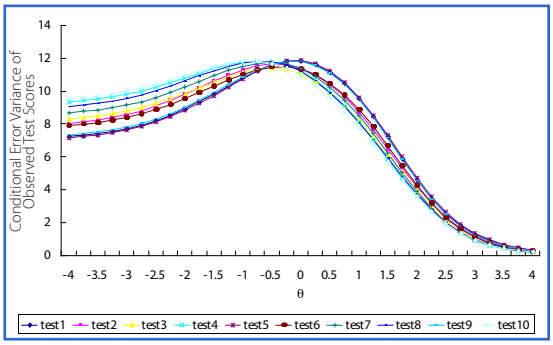


II: With Test Overlap Control

Figure 19: Conditional Error Variance of Observed Test Score (X) for 60-Item Tests Under *Unbalanced* Content Outline *Without* (I) and *With* (II) Test Overlap Control

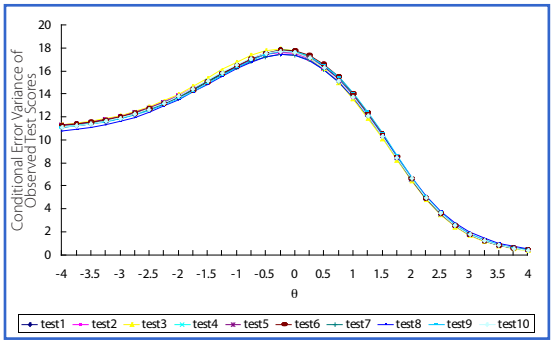


I: Without Test Overlap Control

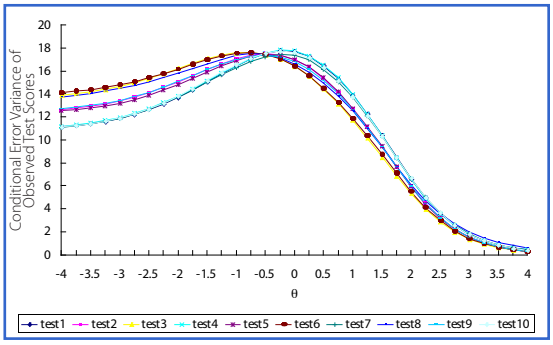


II: With Test Overlap Control

Figure 20: Conditional Error Variance of Observed Test Score (X) for 90-Item Tests Under *Unbalanced* Content Outline *Without* (I) and *With* (II) Test Overlap Control

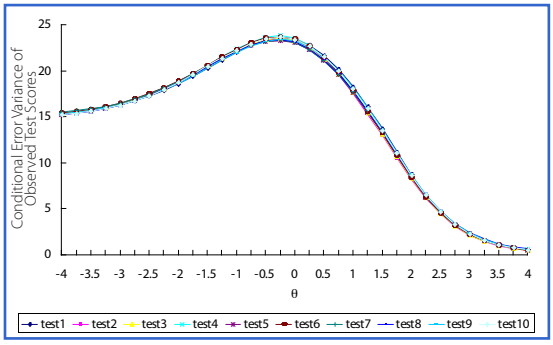


I: Without Test Overlap Control

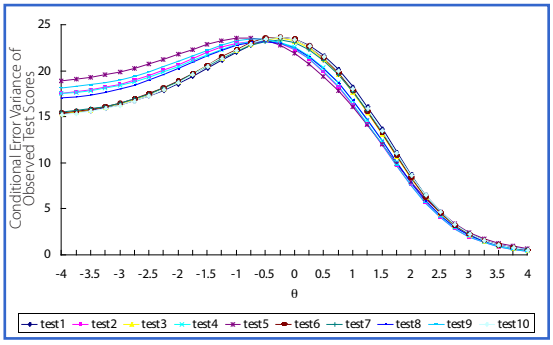


II: With Test Overlap Control

Figure 21: Conditional Error Variance of Observed Test Score (X) for 120-Item Tests Under *Unbalanced* Content Outline *Without* (I) and *With* (II) Test Overlap Control



I: Without Test Overlap Control



II: With Test Overlap Control

To provide further information about differences in the second central moment of $P(\theta)$ and in the CEV of X among the assembled tests, the SD of the second central moment of $P(\theta)$ across the 10 test forms was aggregated across the entire proficiency scale in each experimental condition, and the SD of the CEV of X across the 10 test forms was also aggregated over the entire proficiency scale in each experimental condition (Table 5). Table 5 shows that lower aggregated deviations tend to be associated with tests assembled without exposure control (range from 0.003 to 0.005 for the second central moment of $P(\theta)$; range from 0.079 to 0.162 for the CEV of X), and higher deviations are associated with tests generated with exposure control (range from 0.006 to 0.010 for the second central moment of $P(\theta)$; range from 0.488 to 0.962 for the CEV of X). These findings are consistent to those shown in the previous plots of the second central moment of $P(\theta)$ and the CEV of X.

This finding indicates that constraining test information does not guarantee equivalent levels of test difficulty, variability of item difficulty, and error variance of test score, especially when the test overlap control was imposed in test assembly. This result most likely occurred because a particular test information function may result from various combinations of item difficulties, and the combinations would be more variable with test overlap control. More discussion is given in the Conclusion section.

Table 5: Aggregated Deviations Among the Assembled Tests in TCF, the Second Central Moment of $P(\theta)$, and the CEV of X *With* Exposure Control and *Without* Exposure Control

		TCF		the second central moment of $P(\theta)$		CEV of X	
Content	**Test Length**	**exp.**	**no exp.**	**exp.**	**no exp.**	**exp.**	**no exp.**
Balanced	60	0.016	0.003	0.008	0.005	0.488	0.099
	90	0.017	0.003	0.010	0.004	0.772	0.162
	120	0.016	0.002	0.009	0.003	0.962	0.162
Unbalanced	60	0.017	0.002	0.007	0.003	0.509	0.079
	90	0.015	0.002	0.008	0.004	0.665	0.131
	120	0.012	0.002	0.006	0.003	0.708	0.128

Note: exp. = with exposure control; no exp. = without exposure control

Average Test Overlap Rate

In this study, the value of $\overline{\Omega}_{max}$ under each experimental condition was determined such that it would result in ten test forms that met all assembly constraints and a minimum value for average test overlap rate. The value of $\overline{\Omega}_{max}$ varied with content specification and with the ratio of test length to the size of the item pool. Under the balanced content condition, the values of $\overline{\Omega}_{max}$ were 0.10, 0.12, and 0.18 for 60-, 90-, and 120-item tests, respectively. Under the unbalanced content condition, the values of $\overline{\Omega}_{max}$ were 0.16, 0.22, and 0.34 for 60-, 90-, and 120-item tests, respectively. Specifically, the unbalanced content condition yielded target maximum expected test overlap rate, $\overline{\Omega}_{max}$, larger than the expected baseline test overlap rate, E*(BTOR)*, for 90- and 120-item tests. On the other hand, the balanced content condition generated $\overline{\Omega}_{max}$ smaller than E*(BTOR)* for all test lengths.

Table 6: Summary of Item Usage and Test Overlap Rate for Balanced Content Condition

	Ratio = 0.1 = 60/600		Ratio = 0.15 = 90/600		Ratio = 0.2 = 120/600	
# of Test Forms	**No Control**	**$\overline{\Omega}_{max}$ = 0.10**	**No Control**	**$\overline{\Omega}_{max}$ = 0.12**	**No Control**	**$\overline{\Omega}_{max}$ = 0.18**
0	423 (70.50%)	227 (37.83%)	355 (59.17%)	123 (20.50%)	332 (55.33%)	104 (17.33%)
1	61 (10.17%)	146 (24.33%)	75 (12.50%)	54 (9.00%)	68 (11.33%)	26 (4.33%)
2	29 (4.83%)	227 (37.83%)	47 (7.83%)	423 (70.50%)	35 (5.83%)	236 (39.33%)
3	23 (3.83%)	0 (0.00%)	22 (3.67%)	0 (0.00%)	28 (4.67%)	234 (39.00%)
4	21 (3.50%)	0 (0.00%)	21 (3.50%)	0 (0.00%)	19 (3.17%)	0 (0.00%)
5	6 (1.00%)	0 (0.00%)	16 (2.67%)	0 (0.00%)	19 (3.17%)	0 (0.00%)
6	8 (1.33%)	0 (0.00%)	20 (3.33%)	0 (0.00%)	18 (3.00%)	0 (0.00%)
7	7 (1.17%)	0 (0.00%)	10 (1.67%)	0 (0.00%)	21 (3.50%)	0 (0.00%)
8	7 (1.17%)	0 (0.00%)	10 (1.67%)	0 (0.00%)	16 (2.67%)	0 (0.00%)
9	5 (0.83%)	0 (0.00%)	9 (1.50%)	0 (0.00%)	16 (2.67%)	0 (0.00%)
10	10 (1.67%)	0 (0.00%)	15 (2.50%)	0 (0.00%)	28 (4.67%)	0 (0.00%)
Max Exp	**1**	**0.2**	**1**	**0.2**	**1**	**0.3**
ATOR	0.510	0.084	0.540	0.104	0.633	0.174
E*(BTOR)*	0.100	0.100	0.150	0.150	0.200	0.200
ATOR–E*(BTOR)*	0.410	–0.016	0.390	–0.046	0.433	–0.026

Note: ratio = ratio of test length to item-pool size; Max Exp = maximum item exposure rate.

Comparisons of *ATOR* and E*(BTOR)* are important in evaluating ATA methods because a large difference between *ATOR* and E*(BTOR)* would signal a possible problem in test security resulting from an unacceptable overlap in items among the generated test forms. Tables 6–7 (pages 32 to 33) report *ATOR* and E*(BTOR)* for each experimental condition. Taken as a whole, the results indicate that all no-exposure-control conditions yielded unacceptable test overlap rates. *ATOR* values ranged from 0.510 to 0.633 for the balanced content condition and 0.541 to 0.700 for the unbalanced content condition, and differences between *ATOR* and E*(BTOR)* ranged from 0.390 to 0.433 for the balanced content condition and 0.411 to 0.441 for the unbalanced content condition.

Table 7: Summary of Item Usage and Test Overlap Rate for Unbalanced Content Condition

	Ratio = 0.1 = 60/600		Ratio = 0.15 = 90/600		Ratio = 0.2 = 120/600	
# of Test Forms	**No Control**	**$\overline{\Omega}_{max}$ = 0.10**	**No Control**	**$\overline{\Omega}_{max}$ = 0.12**	**No Control**	**$\overline{\Omega}_{max}$ = 0.18**
0	443 (73.83%)	295 (49.17 %)	392 (65.33%)	239 (39.83%)	361 (60.17%)	230 (38.33%)
1	47 (7.83%)	67 (11.17 %)	60 (10.00%)	73 (12.17%)	52 (8.67%)	74 (12.33%)
2	21 (3.50%)	181 (30.17 %)	29 (4.83%)	114 (19.00%)	30 (5.00%)	55 (9.17%)
3	17 (2.83%)	57 (9.50 %)	14 (2.33%)	97 (16.17%)	23 (3.83%)	53 (8.83%)
4	17 (2.83%)	0 (0.00%)	19 (3.17%)	77 (12.83%)	15 (2.50%)	83 (13.83%)
5	13 (2.17%)	0 (0.00%)	12 (2.00%)	0 (0.00%)	17 (2.83%)	105 (17.50%)
6	12 (2.00%)	0 (0.00%)	12 (2.00%)	0 (0.00%)	14 (2.33%)	0 (0.00%)
7	8 (1.33%)	0 (0.00%)	17 (2.83%)	0 (0.00%)	17 (2.83%)	0 (0.00%)
8	8 (1.33%)	0 (0.00%)	12 (2.00%)	0 (0.00%)	12 (2.00%)	0 (0.00%)
9	5 (0.83%)	0 (0.00%)	13 (2.17%)	0 (0.00%)	15 (2.50%)	0 (0.00%)
10	9 (1.50%)	0 (0.00%)	20 (3.33%)	0 (0.00%)	44 (7.33%)	0 (0.00%)
Max Exp	**1**	**0.2**	**1**	**0.2**	**1**	**0.3**
ATOR	0.541	0.130	0.629	0.214	0.700	0.326
E*(BTOR)*	0.130	0.130	0.194	0.194	0.259	0.259
ATOR– E*(BTOR)*	0.411	0	0.435	0.020	0.441	0.067

Note: ratio = ratio of test length to item-pool size; Max Exp = maximum item exposure rate.

The use of the ordered-item-pooling control procedure to find the item allocation with the smallest average test-overlap rate *(ATOR)* with all test-assembly constraints satisfied reduced, on the average, the value of *ATOR* by 0.44 (i.e., [(0.510 – 0.084) + (0.54 – 0.104) + (0.633 – 0.174)] ÷ 3) for the balanced content condition and by 0.4 (i.e., [(0.541 – 0.13) + (0.629 – 0.214) + (0.7 – 0.326)] ÷ 3) for the balanced content condition. Moreover, acceptable overlap rates were found for all test-length conditions under the balanced content condition because all *ATOR* values were less than the corresponding E*(BTOR)* values. For the unbalanced content condition, acceptable overlap rates were found only for the 60-item condition whereas the 90- and 120-item conditions generated unacceptable overlap rates even with test overlap control. These 90- and 120-item conditions had resultant *ATOR*s greater than their E*(BTOR)*. The reason for this is because the 90- and 120-item conditions had target test overlap rates (or maximum expected test overlap rates, $\overline{\Omega}_{max}$) larger than the corresponding E*(BTOR)* values to begin with, which were 0.22 > 0.194 for the 90-item condition and 0.34 > 0.259 for the 120-item condition.

Distribution of Item Utilization

Table 6 (page 32) shows the distribution of item utilization without exposure control in the 2nd, 4th, and 6th columns, while the distribution of item utilization with exposure control is shown in the 3rd, 5th, and 7th columns for ratio = 0.1, 0.15, and 0.2, respectively, under the balanced content condition. Table 7 (page 33) provides the corresponding results under the unbalanced content condition. Take Ratio = 0.1 under the balanced content condition as an example (Table 6), the second column lists the number of items that appeared on 0, 1, 2, 3, … , 10 forms without exposure control, showing that the number of items that appeared on all ten forms was 10 while the number of items that never appeared on a single form was 423. The third column of Table 6 lists the number of items that appeared on 0, 1, 2, 3, … , 10 forms with exposure control, showing that the use of the ordered-item-pooling control reduced the number of items appearing on all ten forms from 10 to 0 while reducing the number of items never appearing on a single form from 423 to 227. These findings appear to generalize across the test-length (ratio = 0.1 vs. 0.15 vs. 0.2) and content (balanced vs. unbalanced) conditions.

Based on the distribution of item utilization shown in Tables 6–7, this research placed special emphasis on the maximum item exposure rate and the number of unused items to compare the no exposure control and the ordered-item-pooling control procedures. The maximum item exposure rate was computed by dividing the maximum number of forms that an item has appeared on by the number of forms to be assembled (i.e., 10). Tables 6–7 reveal that the maximum item exposure rate and the number

of unused items decrease substantially when the ordered-item-pooling control procedure was implemented. Take ratio = 0.1 under the balanced content condition as an example, when using the ordered-item-pooling control procedure, the maximum item exposure rate reduced from 1 to 0.2, and the number of unused items decreased from 423 to 227. This phenomenon was supported by smaller skewness and range (i.e., maximum-minimum) shown in Tables 8–9, indicating that the item exposure distribution became less skewed and more even when the ordered-item-pooling control was applied. These results generalized across the test length and content conditions.

Table 8: Item Usage (or Exposure) Distribution for the Item Pool Under *Balanced* Content Condition

	Ratio	N	Max	Min	Range	Skewness	Mean	*SD*
Control	0.1	600	0.2	0	0.2	0.000	0.10	0.087
	0.15	600	0.2	0	0.2	–1.150	0.15	0.081
	0.2	600	0.3	0	0.3	–0.872	0.20	0.106
No Control	0.1	600	1.0	0	1.0	2.634	0.10	0.214
	0.15	600	1.0	0	1.0	1.909	0.15	0.256
	0.2	600	1.0	0	1.0	1.452	0.20	0.307

Note: ratio = ratio of test length to item-pool size.

Table 9: Item Usage (or Exposure) Distribution for the Item Pool Under *Unbalanced* Content Condition

	Ratio	N	Max	Min	Range	Skewness	Mean	*SD*
Control	0.1	600	0.3	0	0.3	0.450	0.10	0.108
	0.15	600	0.4	0	0.4	0.389	0.15	0.146
	0.2	600	0.5	0	0.5	0.359	0.20	0.197
No Control	0.1	600	1.0	0	1.0	2.503	0.10	0.221
	0.15	600	1.0	0	1.0	1.904	0.15	0.278
	0.2	600	1.0	0	1.0	1.497	0.20	0.326

Note: ratio = ratio of test length to item-pool size.

Although the ordered-item-pooling control procedure noticeably lowered the maximum item exposure rate, maximum item exposure rates were not acceptable for all conditions. For the balanced content condition, the ordered-item-pooling control procedure yielded acceptable maximum item exposure rates for all test length conditions with r_{max} values no greater than 0.3. For the unbalanced content condition, an acceptable maximum item exposure rate was found for the 60-item condition with r_{max} = 0.3,

whereas unacceptable maximum item exposure rates were found for the 90- and 120-item conditions with r_{max} = 0.4 and 0.5, respectively. This phenomenon indicated that additional control on item exposure rates may be necessary to guarantee an acceptable maximum item exposure rate in the less supportive or more difficult situation for ATA, referring to greater ratios of test length to pool size (i.e., 90/600 = 0.15 and 120/600 = 0.2) and an unbalanced content condition in the study.

Tables 6–7 (pages 32 to 33) also show that when no exposure control was implemented, the unbalanced content condition yielded a larger number of items that appeared on all ten forms than the balanced content condition, and the number of items that appeared on all ten forms increased as the ratio of test length to pool size increased. This occurred because the unbalanced content condition and the larger ratio of test length to pool size represent difficult test-assembly conditions, where the content properties of a target reference form differ from those of the pool and more items are required to be selected into a test form. As a result, a larger percentage of items may be shared by alternate test forms if the item-exposure rate is not controlled under more difficult test-assembly conditions. Specifically, test security breach becomes more serious when test-assembly conditions become more difficult due to the unbalanced content condition and greater ratio of test length to pool size. When the ordered-item-pooling control procedure was implemented, a greater reduction in the number of unused items was observed in the balanced content condition (423 – 227 = 196 in the balanced condition vs. 443 – 295 = 148 in the unbalanced condition for ratio = 0.1). Note that the magnitude of the reduction was similar over all test-length conditions.

Conclusion

Assembling equivalent test forms with minimal test overlap across forms is important in ensuring test security. In ATA, an exposure control method would need to demonstrate its effectiveness by achieving an acceptable average test overlap rate across multiple forms without compromising the conformity to the test-assembly constraints and the test equity of the assembled forms. The ordered-item-pooling control procedure met this standard by first showing that the content constraints were met exactly and the conformity of psychometric constraints was acceptable for all test forms assembled under its test overlap control. However, the degree of test equity in terms of the first central moment of $P(\theta)$ (i.e., TCF), the square root of second central moment of $P(\theta)$, and the conditional error variance of observed test score X (i.e., CEV of X), generated under the ordered-item-pooling control procedure was less than that

yielded under the no test overlap control procedure. The important message from this finding for those resultant psychometric properties is that constraining test information does not guarantee equivalent levels of test difficulty, variability of item difficulty, and error variance of test score, especially when the test overlap control was imposed in test assembly. This result seems reasonable given that a particular test information function could result from various combinations of item difficulties, and the combinations would be more variable with test overlap control. Accordingly, compared to the no exposure control condition, the TCF plots under test overlap control showed greater variation among the assembled tests especially in the lower half of the proficiency scale. Moreover, to ensure test equity in difficulty level, test difficulty may be constrained additionally in the test overlap control condition. However, if test difficulty is added as a test-assembly constraint, other content and psychometric constraints may need to be sacrificed to meet that constraint.

Average test overlap rate is defined as acceptable when the resultant average test overlap rate is smaller than the expected baseline test overlap rate, E*(BTOR)*. The results showed that the average test overlap rates *(ATOR)* were unacceptably high for all tests assembled with the no test overlap control procedure. When the ordered-item-pooling control method was implemented, the average test overlap rate decreased substantially from the no test overlap control condition. This result generalized across test length and content conditions. However, acceptable overlap rates were found only with the tests for the balanced content condition and with the 60-item tests for the unbalanced content condition. The average test overlap rates of the 90-item and 120-item tests generated with the ordered-item-pooling control were still greater than the corresponding E*(BTOR)*.

For the ordered-item-pooling control procedure, a maximum expected test overlap rate (i.e., $\overline{\Omega}_{max}$) is first established. In evaluating the results from the ordered-item-pooling control procedure, it should be noted that this study selects, under each condition, the maximum expected test overlap rate (i.e., $\overline{\Omega}_{max}$) that results when the average test overlap rate (i.e., *ATOR*) is a minimum and ten test forms have met all assembly constraints. The 90- and 120-item tests with the unbalanced content conditions had unacceptable *ATOR*s because they had values larger than the corresponding E*(BTOR)* values to begin with. The high overlap rates for the 90-item and 120-item tests with the unbalanced content specification suggest that the overlap rate would need to be further controlled with some additional constraints in those conditions to ensure adequate test security. However, if stringent control is included as a test-assembly constraint, other content and psychometric constraints may need to be sacrificed to meet that constraint. Moreover, the conditions of a large ratio of test length to item pool

size and test content specification not mirroring the content distribution of the item pool represent poorly supportive situations for automated test assembly. There might be situations in which the ratio of test length to item pool size is so large and the test content specification is so unbalanced that unacceptable test overlap is produced for all ATA approaches with any exposure control technique. The tradeoff between ensuring test parallelism and controlling test overlap rate (and/or item exposure rate) may need to be evaluated by testing specialists and practical criteria for test assembly, administration, and use.

Similarly to the results for average test overlap rate, the maximum item exposure rate equaled 1.0 when test overlap rate control was not implemented whereas the maximum item exposure rate decreased substantially as the ordered-item-pooling control was applied. Additionally, the ordered-item-pooling control procedure yielded fewer numbers of unused items than the no test overlap control procedure. It is particularly noteworthy that greater reduction in maximum item exposure rate and number of unused items was associated with the balanced content condition. Moreover, under the balanced content situation, the maximum item exposure rate decreased from 1 to 0.2 for the 60-item and 90-item tests and from 1 to 0.3 for the 120-item test. However, under the unbalanced content condition, the ordered-item-pooling control procedure yielded acceptable maximum item exposure rate (i.e., 0.2) for the 60-item tests, but not for the 90-item and 120-item tests (i.e., 0.4 and 0.5, respectively). These unacceptable maximum item exposure rates likely also occurred due to higher maximum expected test overlap rates, $\overline{\Omega}_{max}$, and poorly supportive situations for automated test assembly.

Taken together, this study showed that the ordered-item-pooling control procedure appeared to be an effective method in controlling test overlap rate and item exposure rate in most cases. However, it yielded an unacceptable average test overlap rate and maximum item exposure rate in a less supportive situation for ATA, referring to greater ratios of test length to pool size (i.e., 90/600 = 0.15 and 120/600 = 0.2) and the unbalanced content condition in the study. Accordingly, controlling test overlap rate in ATA may not guarantee acceptable maximum item exposure rate when the ratio of test length to pool size is large enough (i.e., 0.15 in the study) under the unbalanced content condition.

The average test overlap rates for the test-length-to-pool-size ratios of 0.15 and 0.2 with the unbalanced content specification were greater than the corresponding E*(BTOR)* as the ordered-item-pooling control was conducted, but the deviations were not substantial (0.02 for the 90-item tests and 0.067 for the 120-item tests). The maximum item exposure rates were unacceptable under these situations. These results indicated that the

ordered-item-pooling control procedure needs to be incorporated with additional constraints to produce acceptable average test overlap rates and maximum item exposure rates in the less supportive contexts for ATA.

The outcomes of the study provide a better understanding of how well the ordered-item-pooling control method functions in automated assembly of multiple forms under various experimental conditions, and are expected to lay the foundation for future research on development of procedures to ensure test security in ATA.

Limitations

When considering the results of this study, several limitations should be kept in mind. First, the generalizability of the results to practical contexts is limited. The content specifications in this study were set up so that complex confounding effects of content specifications and item difficulty would not affect the outcomes. In many realistic test situations, content categories are likely to differ in difficulty. In addition, the number of constraints imposed in some realistic testing situations may be much greater than those imposed in the present study. Simplifying the research situations in this study makes it difficult to generalize the outcomes to more realistic testing situations, such as those in which numerous test-assembly constraints are imposed or there is a complex confounding relationship between content specification and item difficulty.

Second, the statistics for items in the pool used in this study were assumed to be representative of examinees' performance on operational test forms and free from context effects (e.g., subject motivation, item location). More realistic item pools may not fulfill those assumptions when a testing program migrates from paper/pencil to computer-based testing administration modes.

Third, the item pool used in this study was predominantly of medium difficulty and therefore unsuitable for certain testing situations. In contexts involving employment promotion or screening out unqualified candidates, for example, a greater number of items with varying degrees of difficulty may be needed to yield acceptable *ATOR*s. This study did not focus on how well the ordered-item-pooling control method would perform under these conditions.

Finally, the use of ATA techniques represents only part of the complete test assembly process. Tests assembled by ATA techniques still need to be evaluated by testing specialists to ensure that the assembled forms meet desired psychometric, non-psychometric, and practical criteria for test assembly, administration, and use.

Suggestions for Future Research

The most important finding from this study is that the ordered-item-pooling control procedure appeared to be an effective method in controlling test overlap rate and item exposure rate in most cases. However, the study leaves several questions unanswered that could form the basis for future investigations. One essential question to be answered is how well the ordered-item-pooling control procedure performs under different and more realistic conditions. These conditions might include assembling alternate test forms: (a) when item pools have different distributions of item difficulty, (b) when test overlap is completely disallowed or constrained to be very low, (c) when content categories are greater in number and vary in difficulty, and (d) when test speediness and item location effects need to be addressed.

Another important unanswered question is whether the ordered-item-pooling control procedure can be improved to produce acceptable average test overlap rate and maximum item exposure rate in a less supportive situation for ATA. One possible way to address this question is to incorporate the ordered-item-pooling control method with the item exposure control procedure. The new method could be evaluated under the same conditions examined here or under the conditions suggested above.

A third unanswered question is how well the ordered-item-pooling control procedure or this procedure along with item exposure control would perform compared to other exposure control methods (e.g., the *a*-stratified method proposed by Chang and Ying (1999); the Sympson and Hetter (SH) procedure (1985)). Such a study could be conducted under the same conditions examined here or under the conditions suggested above.

Finally, studies might be conducted to assess the effectiveness of the ordered-item-pooling control procedure while imposing different psychometric constraints on test assembly (e.g., a target test information function and a target characteristic function could be imposed to produce alternate forms having identical conditional error variances and conditional true scores).

References

Armstrong, R.D., Jones, D.H., Li, X., & Wu, I.-L. (1996). A study of a network-flow algorithm and a noncorrecting algorithm for test assembly. *Applied Psychological Measurement, 20*, 89–98.

Chang, H.-H., & Ying, Z. (1999). a_stratified multistage Computerized Adaptive Testing. *Applied Psychological Measurement, 23*, 211–222.

Chang, H.-H., & Zhang, J. (2002). Hypergeometric family and item overlap rates in computerized adaptive testing. *Psychometrika, 67*, 387–398.

Chen, S., Ankenmann, R.D., & Spray, J.A. (2003). The relationship between item exposure and test overlap in computerized adaptive testing. *Journal of Educational Measurement, 40*, 129–145.

Chen, S., & Lei, P. (2005). Controlling Item Exposure and Test Overlap in Computerized Adaptive Testing. *Applied Psychological Measurement, 29*, 204–217.

Chen, S., & Lei, P. (2009, in press). Investigating the relationship between item exposure and test overlap: Item sharing and item pooling. *British Journal of Mathematical and Statistical Psychology*.

Chen, S., Lei, P., & Liao, W. (2008). Controlling item exposure and test overlap on the fly in computerized adaptive testing. *British Journal of Mathematical and Statistical Psychology, 61*, 471–492.

Luecht, R.M., & Hirsch, T.M. (1992). Item selection using an average growth approximation of target information functions. *Applied Psychological Measurement, 16*, 41–51.

Lin, C.-J. (2008). Comparisons between Classical Test Theory and Item Response Theory in Automated Assembly of Parallel Test Forms. *Journal of Technology, Learning, and Assessment, 6*(8). Retrieved [date] from http://www.jtla.org.

Luecht, R. M. (1998). Computer-assisted test assembly using optimization heuristics. *Applied Psychological Measurement, 22*, 224–236.

McDonald, R.P. (1999). *Test Theory: a Unified Treatment*. Mahwah, New Jersey: Lawrence Erlbaum Associations, Inc., Publishers.

Parshall, C., Spray, J.A., Kalohn, J., & Davey, T. (2001). *Practical considerations in computer based testing*. Springer-Verlag, N.Y.

Ross, S. M. (1976). *A first Course in Probability*. New York: Macmillan Publishing Inc.

Swanson, L., & Stocking, M.L. (1993). A model and heuristic for solving very large item selection problems. *Applied Psychological Measurement, 17*, 151–166.

Sympson, J.B., & Hetter, R.D. (1985, October). Controlling item-exposure rates in computerized adaptive testing. Proceedings of the 27th Annual Meeting of the Military Testing Association (pp. 973–977). San Diego, CA: Navy Personnel Research and Development Center.

van der Linden, W.J., & Adema, J.J. (1997). Simultaneous assembly of multiple test forms. *Journal of Educational Measurement, 35*, 185–198.

van der Linden, W.J., & Boekkooi-Timminga, E. (1989). A maxmin model for test design with practical constraints. *Psychometrika, 54*, 237–247.

Author Biography

Chuan-Ju Lin is an assistant professor at National University of Tainan, Taiwan, Department of Education. Her current research focuses on measurement issues concerning computer-based testing and group-score assessments. She can be contacted at cjulin@mail.nutn.edu.tw.

J·T·L·A

The Journal of Technology, Learning, and Assessment

www.jtla.org

Technology and Assessment Study Collaborative
Caroline A. & Peter S. Lynch School of Education, Boston College