J·T·L·A

The Journal of Technology, Learning, and Assessment

Volume 6, Number 8 · April 2008

# Comparisons between Classical Test Theory and Item Response Theory in Automated Assembly of Parallel 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 6, Number 8

# Comparisons between Classical Test Theory and Item Response Theory in Automated Assembly of Parallel 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 on-line journal, published by the Technology and Assessment Study Collaborative, Caroline A. & Peter S. Lynch School of Education, Boston College.



**Preferred citation:**

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.

**Abstract:**

The automated assembly of alternate test forms for online delivery provides an alternative to computer-administered, fixed test forms, or computerized-adaptive tests when a testing program migrates from paper/pencil testing to computer-based testing. The weighted deviations model (WDM) heuristic is particularly promising for automated test assembly (ATA) because it is computationally straightforward and produces tests with desired properties under realistic testing conditions. Unfortunately, research into the WDM heuristic has focused exclusively on the Item Response Theory (IRT) methods even though there are situations under which Classical Test Theory (CTT) item statistics are the only data available to test developers.

The purpose of this study was to investigate the degree of parallelism of test forms assembled with the WDM heuristic using both CTT and IRT methods. Alternate forms of a 60-item test were assembled from a pool of 600 items. One CTT and two IRT approaches were used to generate content and psychometric constraints. The three methods were compared in terms of conformity to the test-assembly constraints, average test overlap rate, content parallelism, and statistical parallelism. The results led to a primary conclusion that the CTT approach performed at least as well as the IRT approaches. The possible reasons for the results of the comparability of the three test-assembly approaches were discussed and the suggestions for the future ATA applications were provided in this paper.

J·T·L·A

# Comparisons between Classical Test Theory and Item Response Theory in Automated Assembly of Parallel Test Forms

Chuan-Ju Lin
*National University of Tainan, Taiwan*

## Introduction

Assembling equivalent test forms with minimal test overlap across forms is important in ensuring test security. An ideal goal in alternate test form assembly is to construct test forms that are equivalent in psychometric characteristics, as well as in non-psychometric properties. In practice, however, it may be impossible to achieve such a goal according to the Classical Test Theory (CTT)-based definition of parallel tests, in which true scores and variances of observed test scores across forms must be identical for any possible subpopulation of examinees (Lord & Novick, 1968; Lord, 1980). When CTT conditions for parallelism are not strictly met, post-administration equating and passing score determination are adapted to adjust for differences among test forms so that it makes no difference which form an examinee takes.

The goal of pre-equating is to derive equating transformations before a test is administered intact. Automated assembly of multiple forms can result in pre-equated forms if conditions for parallelism are met. That is, if the definition of test parallelism holds for the constructed forms, test forms are indistinguishable to the examinee and test scores of different test forms are comparable to each other. Therefore, the goal of pre-equating is considered to be achieved. Assembling multiple test forms prior to administration is an appealing idea because, theoretically, if pre-equated test forms are truly parallel and maintain minimal item overlap or duplication, post-administration corrections will not be required. This result holds because the differences in an examinee's scores on alternate test forms should occur from random fluctuation rather than systematic differences in the test forms. When designed properly, pre-equated (parallel) test forms could be administered to candidates in high-stakes testing situations with a nominal amount of post-testing delay in reporting scores.

The forms would be equitable in terms of fairness to candidates, and test security problems would be manageable. As computer technology becomes more prevalent, pre-equated parallel test forms can be obtained efficiently via automated test assembly or ATA procedures.

Swanson and Stocking (1993) first proposed the use of the WDM heuristic for automated test assembly. Stocking, Swanson, and Pearlman (1993) subsequently applied the WDM heuristic to the automated assembly of fixed forms using real item pools and showed that it produced the best possible tests given the nature of the item pools and constraints imposed. In these previous studies, IRT methods were used exclusively to calibrate item response data and generate psychometric constraints. However, there may be situations under which classical item statistics are the only data available. Not all testing programs can use IRT methods to calibrate item response data because 0/1 item response data may be unavailable, and the test assembler may have to rely on classical item statistics alone in creating parallel forms. For example, a testing company may be hired to create parallel forms from an existing pool for which only $p$-values and $r_{pbis}$ values are available. Under these conditions, test assembly methods using classical item statistics would provide a possible solution to the test assembly problem. Accordingly, it was important to investigate how a Classical Test Theory (CTT) approach might affect the automated assembly of alternate test forms.

## Research Question

The *primary question* addressed in this study was, "Do test assembly methods using classical item statistics yield test forms that are as parallel as those constructed using IRT methods?" Such comparisons can be made by contrasting the classical item-statistics approach with conditional IRT approaches in producing alternate test forms using the degree of congruence among the distributions of item response functions of alternate test forms as an index of parallelism, which is described in detail in later sections.

Although the construction of alternate test forms with classical statistics and IRT-based functions is not based on this strict definition of parallelism, it seems reasonable to presume that the approaches based on IRT would yield test forms of higher degrees of parallelism. This might occur because, by definition, an item characteristic function is a regression function on the examinees' unidimensional latent trait, $\theta$. In other words, by nature, IRT models deal with item difficulty and variability issues conditionally and thereby are more sensitive to individual item characteristics than the classical approach is. However, it may be the case that IRT-based

methods do not produce better parallelism due to factors related to the algorithm used for automated test assembly. For example, when the number of the IRT-based constraints (e.g., 33 constraints, or TTIF at 33 θ values) is much greater than that of the CTT-based constraints (i.e., 2 constraints, or expected observed-score mean and standard deviation) for automated test assembly, the IRT-based methods may produce less optimal tests, and thereby less parallel ones.

## Purpose of Study

The purpose of this study was to investigate the degree of parallelism of test forms constructed with the WDM heuristic (Swanson & Stocking, 1993) using both classical and IRT approaches. The criteria used to evaluate test parallelism derived from these ATA approaches were conformity to the ATA constraints, test overlap rate, content parallelism, and statistical parallelism. More specifically, the study focused on how these test properties were affected by the ATA approach used (CTT vs. IRT) and its statistical constraints.

The major index of statistical parallelism used in this study was based on the parallelism definition developed by van der Linden and Luecht (1998), which is that for two tests to be exactly parallel, they must have identical moments of $P(\theta)$ for all values of θ, and thus they have identical distributions of $P(\theta)$ for all values of θ. This index was adapted to evaluate test parallelism resulting from each ATA approach because it could provide a more strict standard to evaluate which method (either classical-related or IRT-related) generates multiple test forms with the highest degree of parallelism. However, the strongly parallel index may be an overly strict criterion for statistical parallelism, and thereby an alternative index, Smirnov statistic $T$ (van der Linden & Luecht, 1998; Conover, 1980), based on a less stringent criterion also was examined.

Even though the definition of parallelism involves all $m$ moments, in a practical sense, the important moments to examine between the reference test and the constructed tests are the first moment and the second central moment. The first moment is critical in assessing the form difficulty or conditional difficulty of the constructed tests at certain values of θ, and the second central moment is important in the evaluation of the form variability or variability of $P(\theta)$ at certain values of θ. Accordingly, observed test mean and standard deviation, the first moment and standard deviation of $P(\theta)$, and the conditional error variance of observed test score X (*CEV* of *X*) also were examined in this study to evaluate the degree of statistical parallelism.

# Literature Review

## Automated Test Assembly

Automated test assembly has gained much attention in the measurement community in recent years due to faster computer processing and larger item pools for assembling tests (e.g., Ackerman, 1989; Boekkooi-Timminga, 1990; Swanson & Stocking, 1993; van der Linden, 1987, 1998a, 1998b). The methods of automated test assembly provide efficient test construction and may ensure conformity to specified criteria or targets in test construction such as the equivalence of test forms based on a certain test blueprint. An item pool stored on a computer may contain hundreds or even thousands of items that have various psychometric (e.g., conventional item *p*-values, point-biserial correlation coefficients, and IRT-based a-, b-, and c-parameter estimates) and non-psychometric (e.g., content categories) attributes. These attributes can be used to automatically assemble test forms by modeling desired test specifications using specially designed computer software.

Test forms should be assembled to meet specifications for both psychometric and non-psychometric properties of the tests. Psychometric attributes of items may refer to classical item statistics, IRT-based item parameter estimates (i.e., a-, b-, and c- parameter estimates), item-response functions, or item information functions. The test-level psychometric properties are often functions of the item attributes (Luecht, 1998). For example, the test information function equals the sum of the item information functions, and the test mean equals the sum of the item difficulties (i.e., *p*-values). Non-psychometric specifications for test assembly refer to attributes that are not related to statistical characteristics of tests and include factors such as test length, test content, number of test forms to be constructed, item format, item sets, item enemies (or item exclusion), and item-exposure rate (or item usage frequency).

The basic elements required in implementing automated methods for assembling parallel tests are usually (a) the test length, (b) content constraints (c) desired psychometric properties, and (d) number of test forms to be assembled. Weights for each of these test specifications or constraints are sometimes required. The weights allow for some constraints to be emphasized over others in most of the automated test assembly methods. In the Weighted Deviations Model approach (Swanson & Stocking, 1993), the weights also serve the purpose of placing all constraints that are evaluated on different metrics on equal footing.

Due to recent advances in ATA procedures and faster computer processing, the automated assembly of alternate test forms for online delivery

provides an alternative to computerized adaptive testing (CAT) when a testing program migrates from traditional paper/pencil (p/p) testing to computer based testing (CBT). The automated assembly of alternate test forms for online delivery may be less problematic in comparability than CATs because the former testing format (e.g., fixed forms – same test length and set of items) may be more similar to the p/p format. Additionally, scoring for ATA procedures can remain on a percent-correct scale across all multiple forms – a score scale that is understood by most examinees (or candidates) and program directors. Furthermore, different test-taking behaviors or strategies are not required of candidates so that lengthy explanations or tutorials of test-taking procedures may not be needed for ATA procedures, even when the tests are computer based. Finally, item development need not be sensitive or responsive to particular levels of candidate proficiency because items are not tailored specifically to each examinee's skill level. In most CAT environments, for example, item pools must either be rich in items that measure well across all proficiency levels, or near the passing score. Although the latter would also be the goal of a good fixed-length test, a lack of good items at particular levels of proficiency would not preclude the test developer from constructing multiple test forms using ATA methods.

## Automated Assembly of Alternate Test Forms Based on CTT

Automatically assembled and pre-equated parallel test forms can be developed either within the framework of CTT or item response theory (i.e., IRT). Automated assembly of alternate test forms based on the CTT definition of strictly parallel forms is unrealistic because it is difficult to develop software consistent with this definition (i.e., first- and second- order equities across forms hold for all possible subpopulations). Consequently, in many studies of automatically constructing parallel forms with classical item statistics, identical overall observed-score means and variances for all examinees are commonly defined goals (e.g., Armstrong, Jones, & Wang, 1994; Gibson & Weiner, 1998). The CTT-based statistical indices are easy to compute, manipulate, and understand by lay persons, but they will vary from sample to sample without extensive pre-testing.

## Automated Assembly of Alternate Test Forms Based on IRT

Alternatively, within the framework of IRT and automated test construction, equivalent 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., Luecht, 1998; 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). However, Samejima's definition of parallelism does not necessarily yield identical observed-score distributions because test information functions are only related to the asymptotic error variance of proficiency estimates on the θ-scale rather than the true-score distribution.

van der Linden and Luecht (1998) proposed an IRT-based method for constructing strongly parallel tests by matching items on item response functions. They noted that "Test forms with pairwise identical response functions have equal true scores and observed-score variances for each examinee in the population for which the IRT model holds and are therefore parallel" (van der Linden & Luecht, 1998, p. 402). Thus, the IRT definition of strongly parallel forms will guarantee the equivalence of the observed-score distributions within the CTT strict definition of parallelism because item response functions, test characteristic functions (i.e., true scores), and test information functions (i.e., error variances) are all identical across forms. Since the IRT definition of strongly parallel forms refers to equivalence in item response functions across forms, the criterion that parallel test forms have identical distributions of item response functions conditional on θ is an appropriate standard to evaluate the degree of parallelism of alternate test forms (van der Linden & Luecht, 1998).

With the IRT definition of strong parallelism, if the item response function, $P_i(\theta)$, represents the conditional difficulty of the $i^{th}$ item for a person with latent trait (θ), then parallel test forms would refer to test forms that satisfy the requirement that the distribution of $P_i(\theta)$ is the same for each θ value across test forms, and thus the degree of congruence in the distributions of $P_i(\theta)$ across forms can be used to indicate the degree of test parallelism. This definition of strongly parallel test forms is in line with one proposed by McDonald (1999). McDonald defined two test forms as *item-parallel* if they consist of paired items with identical item parameters.

Nevertheless, this requirement may be too stringent for automated test assembly. McDonald (1999) proposed definitions for other degrees of test parallelism between test forms, one of which is that test forms are regarded as TCC-parallel if they have identical test characteristic curves or functions. Assembling alternate test forms by matching a target test characteristic curve is of practical value when equivalence of test difficulty, true score, or passing score for alternate test forms is the major concern for testing programs. Accordingly, a much more relaxed requirement in which the first moment of conditional difficulty (the test characteristic function or TCF) is identical across test forms can be considered and used

as a statistical constraint for automated assembly of parallel test forms. The test characteristic curve may be anchored by fixing two points on the curve if the points are chosen from the area surrounding the inflection point of the test characteristic curve. If these two corresponding values of θ mark critical points on the latent trait metric, a requirement of identical first moment of conditional difficulty at these two points across test forms might be sufficient to assemble test forms with equivalent test characteristic curves.

### Strict Standard of Evaluating Statistical Parallelism

Not all of the methods for assembling parallel tests discussed here are guaranteed to produce truly parallel test forms. Even though multiple test forms have identical classical statistical properties (e.g., means or standard deviations of observed scale distributions), identical test information functions (TIFs), or identical test characteristic functions (TCFs), they may not match the criteria described in the stringent or strong definitions of parallel test forms based on CTT and IRT. The definition of *item parallelism* proposed by van der Linden and Luecht (1998) and McDonald (1999) is a more stringent constraint for construction of equivalent test forms than are those of *classical-related parallelism* (e.g., equivalent means and standard deviations of observed scale distributions), *TCF parallelism* or *TIF parallelism*. Accordingly, the degree of the congruence in the distributions of item response functions conditional on θ (i.e., $P_i(\theta)$) across forms provides a more strict standard to evaluate which method (either classical-related or IRT-related) yields alternate test forms with the highest degree of parallelism.

## Methods

The test-assembly methods compared in this study are introduced in the following section. Afterwards, the characteristics of the item pool, the properties of the reference form on which to base the test-assembly constraints, and the criteria used to evaluate test parallelism are presented.

### Test-assembly approach

The primary independent variable in the study was test-assembly approach. Three approaches were compared: Classical Test Theory (CTT) with a target test mean and a target test standard deviation, Item Response Theory (IRT) with a target test information function (IRT-TTIF), and Item Response Theory with a target test characteristic function at two values of θ (IRT-TTCF2P). All three approaches were implemented using the weighted deviations model (WDM) heuristic.

## WDM Heuristic

The algorithm used for automated test construction for this study was the weighted deviations model (WDM) heuristic described 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 a greedy algorithm is to maximize a set function by choosing successively the next element that yields the greatest improvement in some criterion value if that element exists. That is, the goal of the greedy heuristic is the pursuit of the maximum improvement at each iteration of some procedure to achieve monotonically better progress in approaching the optimal solution. Within the context of test assembly, a greedy heuristic refers to the rule of item selection by which items are selected sequentially so that those chosen first provide the best improvement when conforming to all the constraints simultaneously. Heuristic algorithms are designed to quickly find the best possible solution to the item selection problem rather than to seek an optimal solution in the sense of exactly meeting all constraints simultaneously. The heuristics may yield appropriate solutions in the sense that the item pool cannot fulfill the target-test constraints perfectly or provide ideal tests.

The WDM heuristic was selected as the automated test assembly method primarily because of its computational simplicity and its flexibility in handling both IRT and CTT item statistics. The method does not require a sophisticated mathematical background to understand or to implement. Additionally, the method always produces the best possible test given the nature of the item pools and constraints imposed rather than seeking an optimal test in the sense of exactly meeting all constraints simultaneously. This is an advantage over many 0–1 linear programming (LP) models when the item pool cannot fulfill the target-test constraints perfectly because the 0–1 linear programming approach still seeks an optimal test and often results in the problem of infeasibility. That is, when the number of constraints (or test specifications) is very large and the item pool cannot supply the required number of qualified items, the test cannot be constructed with the 0–1 linear programming approach.

Implementation of the WDM method typically requires the test constructor to specify: (a) a set of content constraints (i.e., the test-content blueprint or outline), (b) the test length, (c) the statistical properties of the tests to be constructed (i.e., the target or reference-form requirements), and (d) the number of test forms to be drawn. In addition, the test constructor must provide weights for each of the constraints. The weights serve a dual purpose. First, they allow for some of the constraints to be emphasized over others. Second, constraints that are naturally evaluated on different metrics can be placed on equal footing with all other

constraints. This redistribution of the importance of constraint factors is necessary because the WDM approach is based on the evaluation of each item relative to the distance of the (positive) deviations of its content and statistical properties from those required on the target or reference form.

## Characteristics of the Item Pool

### The Known or True Item Pool

The item pool used in this study was derived from ten forms of a particular Mathematics Usage Test, and covered four content areas: 152 content-A items, 127 content-B items, 147 content-C items, and 174 content-D items. This mathematics test is a 60-item exam that measures mathematics achievement in college-bound high school students. The ten forms of the Mathematics Test were calibrated separately using the 3-parameter logistic model (3-PLM) in the computer program, *Bilog-MG* (Zimowski, Muraki, Mislevy & Bock, 2003). The calibrations were not scaled or linked to a common metric because the data for the ten forms were collected from ten randomly equivalent groups with 2000 examinees for each group, and the purpose of obtaining the item-parameter estimates was simply to have a pool of item parameters on which to base the study. Each calibration was performed using the same prior latent trait density, and the default prior in the *Bilog-MG* program is $\theta \sim N(0,1)$. The average item parameters,

$$\bar{a}, \bar{b} \text{ and } \bar{c},$$

in the pool were 0.991, 0.112, and 0.171, respectively. Additionally, the average item *p*-values and point-biserial correlations in the pool were 0.6 and 0.41, respectively.

Because item pools in practice were based on estimates of item characteristics rather than known characteristics, additional steps were undertaken to create the most realistic and appropriate item and test indices for the three test assembly approaches used in this study (CTT, IRT-TTIF, and IRT-TTCF2P). These steps are described below.

### CTT Data

The 600 items from the item pool with known item parameters were used to generate 20000 simulees with dichotomously-scored (i.e. 0/1) data through a simulation by assuming that the proficiency levels θs of examinees are normally distributed with mean of 0 and standard deviation of 1 (i.e., N(0,1)) and that the probability of a correct response to any item, $P(\theta)$, fits the 3-PLM. The generated 0/1 responses were then used to calculate the conventional item difficulty (*p*-value) and item discrimination (point-biserial correlation coefficient or $r_{pbis}$) for each item in the pool. The

resultant average item $p$-values and $r_{pbis}$ indices in the pool were 0.559, and 0.395, respectively.

### IRT Data

The same 0/1 responses derived from the simulation above were then calibrated using the computer program, *Bilog-MG*, in a single computer run. This run produced item parameter estimates for the 600-item pool

$$(\hat{a}_i, \hat{b}_i, \hat{c}_i, \text{i}=1, \ldots, 600)$$

that were used in implementing the two IRT-based automated test assembly methods examined in this study (IRT-TTIF and IRT-TTCF2P). The average item parameter estimates in the pool were 0.965, 0.127, and 0.178, respectively.

## Test-Content Outline

The original content areas from the Mathematics Usage Test were not used in this study to eliminate the likely confounded relationship between test content area and item difficulty. Content area and item difficulty confounding was reduced to provide fairer comparisons among test assembly approaches. To eliminate this possible source of confounding, one of four possible content classifications (A, B, C, or D), assigned to an arbitrary distribution, was randomly selected and assigned to each item in the pool. From this process, the item pool of 600 items resulted in 152 content-A items, 127 content-B items, 147 content-C items, and 174 content-D items (Table 1, next page).

In real testing situations, content outlines need not be the same as the content distributions of the item pool. For many professional certification or licensure programs, item pools frequently have content distributions that differ from the test content outline. For example, there may be fewer items in one content area than another because those items are more difficult to create. And thereby we specified, in this study, a hypothetical reference content outline to represent a content framework that did not mirror the content distribution of the item pool. The test-content outline is presented in Table 1.

**Table 1: Content Distribution of the Item Pool and Test-content Outline**

| Content Area | Item Pool | Test |
|---|---|---|
| A | 152 (25%) | 9 (15%) |
| B | 127 (21%) | 3 (5%) |
| C | 147 (25%) | 18 (30%) |
| D | 174 (29%) | 30 (50%) |

## Reference Test Distribution

The reference test distribution was created from items in the pool. To ensure sufficient items for test assembly, the reference distribution was chosen to mirror the item pool in terms of the test information function (i.e., TIF) and its TIF was peaked over the middle part of the proficiency range (i.e., where the TIF for the item pool was peaked). The psychometric properties from the reference distribution were used as statistical constraints for automated test assembly. The psychometric attributes in a reference distribution were expected to affect the comparability between the CTT and the IRT approaches. If the psychometric attributes in a reference distribution are very different from those in the pool, there may be insufficient items to match the specified constraints for test construction no matter what method is used. The goal of this study is to identify the variables other than pool characteristics (e.g., sufficient items or not) that affect the comparability between the CTT and IRT approaches, and thereby the reference distribution mirroring the item pool was specified for this study to ensure sufficient items for test assembly.

The combination of content outline and reference test distribution specified previously yielded a reference test. In this study, the length of the reference test was 60 items. The psychometric properties were specified under each test-assembly approach. Three approaches to establishing the statistical constraints were examined here. Under the classical test theory (CTT) approach, the expected observed test score or form difficulty and expected standard deviation of observed test scores were used as the statistical constraints. Under the IRT-TTIF approach, a target test information function was specified. Under the IRT-TTCF2P approach, the target test characteristic function at two values of $\theta$, or target conditional test difficulty at two values of $\theta$ was specified. None of these constraint sets is guaranteed to produce truly parallel test forms, as will be revealed later. The psychometric properties from the reference distribution were described in detail in the section to follow.

### Reference Form Difficulty

The difficulty of the reference test was computed from the 60 item parameters of a reference test by assuming that the latent trait distribution, $g(\theta)$, was N(0,1). The difficulty of a given item (i.e., $p$-value or $p_i$) on the reference test equaled

$$p_i = \int_{-\infty}^{+\infty} P_i(\theta) g(\theta) d(\theta),$$

where $P_i(\theta)$ is the item response function for a given item $i$ based on the usual 3-PLM mentioned previously. The difficulty of the reference form equaled

$$\sum_{i=1}^{60} p_i$$

where the sum is over the 60 items on the reference form, and this value was 33.077.

### Expected Observed Score Distribution

The expected observed-score distribution was obtained from the 60 item parameters of a reference test by assuming that $\theta \sim N(0,1)$ and using the recursive procedure described by Lord and Wingersky (1984). This computation was used to obtain the expected observed-score variance, VAR(X), on the reference form, and ultimately, the observed-score standard deviation, $\{VAR(X)\}^{1/2}$ or $SD_X = 11.541$.

### Target Test Information Function

The 60 item parameters from a reference form were used to calculate a target test information function, or $\Sigma[I_i(\theta)]$, where the sum was across items. The target test information function was calculated from the reference form at 33 values of $\theta$, ranging from –4.00 to +4.00 in increments of 0.25.

### Target Conditional Test Characteristic Function on Two Discrete Points

The 60 item parameters from a reference form also were used to calculate a target test characteristic function, or $\Sigma[P(\theta)]$, where the sum was across items. The target test characteristic function was calculated from a reference form for two values of $\theta$, –1.5 and 1.5. These two values of $\theta$ were chosen so that the target test characteristic curve could be anchored by fixing two points on the curve. Two critical values on the $\theta$ scale were expected to be sufficient to anchor or define the test characteristic curve.

Only two values of θ were taken because there is a fairly well-known phenomenon that occurs when the test characteristic function, $\Sigma[P(\theta)]$, is used as a statistical constraint in ATA problems. Because individual item characteristic functions (ICCs) are not additive, their sum is not logistic. Hence, there is a built-in bias in the use of $\Sigma[P(\theta)]$ as a constraint and the constructed tests do not meet specifications.

## ATA Constraints

The ATA problem was to assemble six tests with each of the CCT, IRT-TTIF, and IRT-TTCF2P approaches that followed the content and psychometric constraints specified from the reference form. Accordingly, all assembled test forms needed to be 60 items in length, and follow the content distribution or outline given in Table 1 (page 14). In terms of psychometric properties, each of the six tests assembled with the CTT method needed to have a difficulty of 33.077, and an observed-score standard deviation of 11.541, each of those assembled with the IRT-TTIF method was constrained to have the test information function matching the target function from the reference form, and each of those assembled with the IRT-TTCF2P approach was required to have the test characteristic function matching the target function from the reference form at two values of θ.

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, the actual psychometric and content constraints were arbitrarily defined in Table 2 through Table 4, and all constraints were weighted equally.

**Table 2: CTT Constraints**

| | | |
|---|---|---|
| 1 | Test Difficulty: | $32.577 \leq \Sigma(p\text{-values}) \leq 33.577$ |
| 2 | Observed-score variability: | $11.041 \leq SD_X = \Sigma\{p(1-p)\}^{1/2} r \leq 12.041$ |
| 3 | Content area A: | 9 items |
| 4 | Content area B: | 3 items |
| 5 | Content area C: | 18 items |
| 6 | Content area D: | 30 items |

## Table 3: IRT-TTIF Constraints

| # of Constraint | Lower Bound | TIF (θ) | Upper Bound | TTIF from Reference |
|---|---|---|---|---|
| 1 | 0.049 | ≤ TIF(θ = –4.00) ≤ | 0.133 | 0.091 |
| 2 | 0.117 | ≤ TIF(θ = –3.75) ≤ | 0.197 | 0.157 |
| 3 | 0.227 | ≤ TIF(θ = –3.50) ≤ | 0.303 | 0.265 |
| 4 | 0.402 | ≤ TIF(θ = –3.25) ≤ | 0.474 | 0.438 |
| 5 | 0.669 | ≤ TIF(θ = –3.00) ≤ | 0.737 | 0.703 |
| 6 | 1.057 | ≤ TIF(θ = –2.75) ≤ | 1.121 | 1.089 |
| 7 | 1.596 | ≤ TIF(θ = –2.50) ≤ | 1.656 | 1.626 |
| 8 | 2.311 | ≤ TIF(θ = –2.25) ≤ | 2.367 | 2.339 |
| 9 | 3.211 | ≤ TIF(θ = –2.00) ≤ | 3.273 | 3.242 |
| 10 | 4.338 | ≤ TIF(θ = –1.75) ≤ | 4.386 | 4.362 |
| 11 | 5.672 | ≤ TIF(θ = –1.50) ≤ | 5.716 | 5.694 |
| 12 | 7.219 | ≤ TIF(θ = –1.25) ≤ | 7.259 | 7.239 |
| 13 | 8.955 | ≤ TIF(θ = –1.00) ≤ | 8.991 | 8.973 |
| 14 | 10.827 | ≤ TIF(θ = –0.75) ≤ | 10.859 | 10.843 |
| 15 | 12.765 | ≤ TIF(θ = –0.50) ≤ | 12.793 | 12.779 |
| 16 | 14.671 | ≤ TIF(θ = –0.25) ≤ | 14.695 | 14.683 |
| 17 | 16.414 | ≤ TIF(θ = 0.00) ≤ | 16.434 | 16.424 |
| 18 | 17.846 | ≤ TIF(θ = 0.25) ≤ | 17.862 | 17.854 |
| 19 | 18.865 | ≤ TIF(θ = 0.50) ≤ | 18.877 | 18.871 |
| 20 | 19.437 | ≤ TIF(θ = 0.75) ≤ | 10.445 | 19.441 |
| 21 | 19.495 | ≤ TIF(θ = 1.00) ≤ | 19.499 | 19.497 |
| 22 | 18.857 | ≤ TIF(θ = 1.25) ≤ | 18.865 | 18.861 |
| 23 | 17.371 | ≤ TIF(θ = 1.50) ≤ | 17.383 | 17.377 |
| 24 | 15.112 | ≤ TIF(θ = 1.75) ≤ | 15.128 | 15.120 |
| 25 | 12.375 | ≤ TIF(θ = 2.00) ≤ | 12.395 | 12.385 |
| 26 | 9.507 | ≤ TIF(θ = 2.25) ≤ | 9.531 | 9.519 |
| 27 | 6.847 | ≤ TIF(θ = 2.50) ≤ | 6.875 | 6.861 |
| 28 | 4.658 | ≤ TIF(θ = 2.75) ≤ | 4.690 | 4.674 |
| 29 | 3.037 | ≤ TIF(θ = 3.00) ≤ | 3.073 | 3.055 |
| 30 | 1.927 | ≤ TIF(θ = 3.25) ≤ | 1.967 | 1.947 |
| 31 | 1.202 | ≤ TIF(θ = 3.50) ≤ | 1.246 | 1.224 |
| 32 | 0.743 | ≤ TIF(θ = 3.75) ≤ | 0.791 | 0.767 |
| 33 | 0.454 | ≤ TIF(θ = 4.00) ≤ | 0.506 | 0.480 |
| 34 | Content area A: | 9 items | | |
| 35 | Content area B: | 3 items | | |
| 36 | Content area C: | 18 items | | |
| 37 | Content area D: | 30 items | | |

**Table 4: IRT-TTCF2P Constraints**

| Constraint | | Lower Bound | TTCF (θ) | Upper Bound | TTCF from Reference |
|---|---|---|---|---|---|
| 1 | Conditional Difficulty at θ = -1.5: | 16.636 | ≤ $\Sigma[P_i(\theta = -1.5)$ ≤ | 17.636 | 17.136 |
| 2 | Conditional Difficulty at θ = 1.5: | 50.656 | ≤ $\Sigma[P_i(\theta = +1.5)$ ≤ | 51.656 | 51.156 |
| 3 | Content area A: 9 items | | | | |
| 4 | Content area B: 3 items | | | | |
| 5 | Content area C: 18 items | | | | |
| 6 | Content area D: 30 items | | | | |

## Criteria Used to Evaluate Test Parallelism

### Average Test Overlap Rate (ATOR)

Test overlap rate is an important index to consider in ensuring the security of test items. In evaluating the test parallelism produced by the three ATA methods, it should be emphasized that test overlap rates were simply observed as an outcome variable and used as a supplemental criterion to evaluate test parallelism in the study. For example, when all ATA approaches yield similar degrees of test parallelism, the method with smaller *ATOR* can be considered as better than the others in assembling parallel forms because it generates a comparable degree of test parallelism that would be less likely to compromise test security. Test overlap rate refers to the proportion (or percentage) of items shared by a pair of constructed 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 pairwise constructed forms, and then taking the average over all of these percentages. To derive this index, a random variable $Y$ is defined as the number of common items shared by any paired tests, and $Y/n$ is the overlap rate for any two tests, where $n$ is the test length and $Y = 1, 2, 3, ..,y,.., n$. $Z$ is defined as the number of all possible paired test forms, given that $C$ test forms are constructed, and equals

$$\binom{C}{2}.$$

Accordingly, the average test overlap rate (*ATOR*) is defined mathematically as follows.

$$ATOR = \frac{\sum \frac{Y}{n}}{Z}, \text{ where } \sum \frac{Y}{n}$$

is the sum of test overlap rates for all possible paired tests.

In this study, *ATOR* was computed and compared to an expected baseline item overlap rate, *E*(*BOR*) (Chen, Ankenmann, & Spray, 1999). *E*(*BOR*) is the test-overlap rate when only the content constraints are imposed for automated test assembly. The value of *E*(*BOR*) could be used as a benchmark for constraining test overlap rate that would be less likely to compromise test quality because this index ensures that content specifications will be met. Accordingly, comparisons of *ATOR* and *E*(*BOR*) are important in evaluating test equity yielded by the ATA methods because a large difference between *ATOR* and *E*(*BOR*) would signal a possible problem in test security resulting from unacceptable overlap in items among the generated test forms. When a large difference occurs between *ATOR* and *E*(*BOR*), the item-exposure rate may need to be specified to reduce the item-reuse frequency for constructed test forms.

## Index of Content Parallelism (CP)

A constructed test is parallel to the reference test in content if it has the same content distribution as the reference test. A measure of this similarity is the percentage of content specifications met. For example, a constructed test that has a content distribution of 8 A items, 4 B items, 18 C items, and 30 D items has content parallelism of 58/60×100% = 97%, given that the reference test has a content distribution of 9 A items, 3 B items, 18 C items, and 30 D items.

## Index of Statistical Parallelism

The degree of statistical parallelism was evaluated using the indices of the degree of distributional congruence (i.e., Item characteristic curve parallelism or *ICCP*), Smirnov statistic *T*, test difficulty and variability indices, and the conditional error variance of observed test score X (i.e., *CEV* of *X*). These indices were introduced in detail below.

### *Item characteristic curve parallelism (ICCP)*

According to van der Linden and Luecht (1998), a constructed test form is strongly parallel to the reference form if and only if $\Sigma[P_i(\theta)]^m = \Sigma[P^*_i(\theta)]^m$, $m = 1, 2,\ldots, n$, where $P_i(\theta)$ is the item characteristic curve for the ith item on the reference test, $P^*_i(\theta)$ is the true item characteristic curve for the $i^{th}$

item on the constructed test, $m$ is the $m^{th}$ (noncentral) moment, and $n$ is the number of items in a test. In other words for two tests to be exactly parallel, they must have identical moments of $P(\theta)$ for all values of $\theta$, and thus they have identical distributions of $P(\theta)$ for all values of $\theta$.

The approach used in this study to describe the commonality between the two conditional distributions of $P(\theta)$ was a measure of the percent of overlap of the two distributions. This overlap percentage has been used in other studies to describe the degree to which two distributions are congruent (Spray & Miller, 1992). Based on the equation developed by Spray and Miller (1992), the degree to which the distributions at a particular latent-ability level (for the reference test and the constructed test) are congruent has been defined as $OVR(\theta)100\%$, where

**Equation 1:**

$$OVR(\theta) = \sum_{i=1}^{k} MIN[f\{P(\theta)\}, f\{P^*(\theta)\}],$$

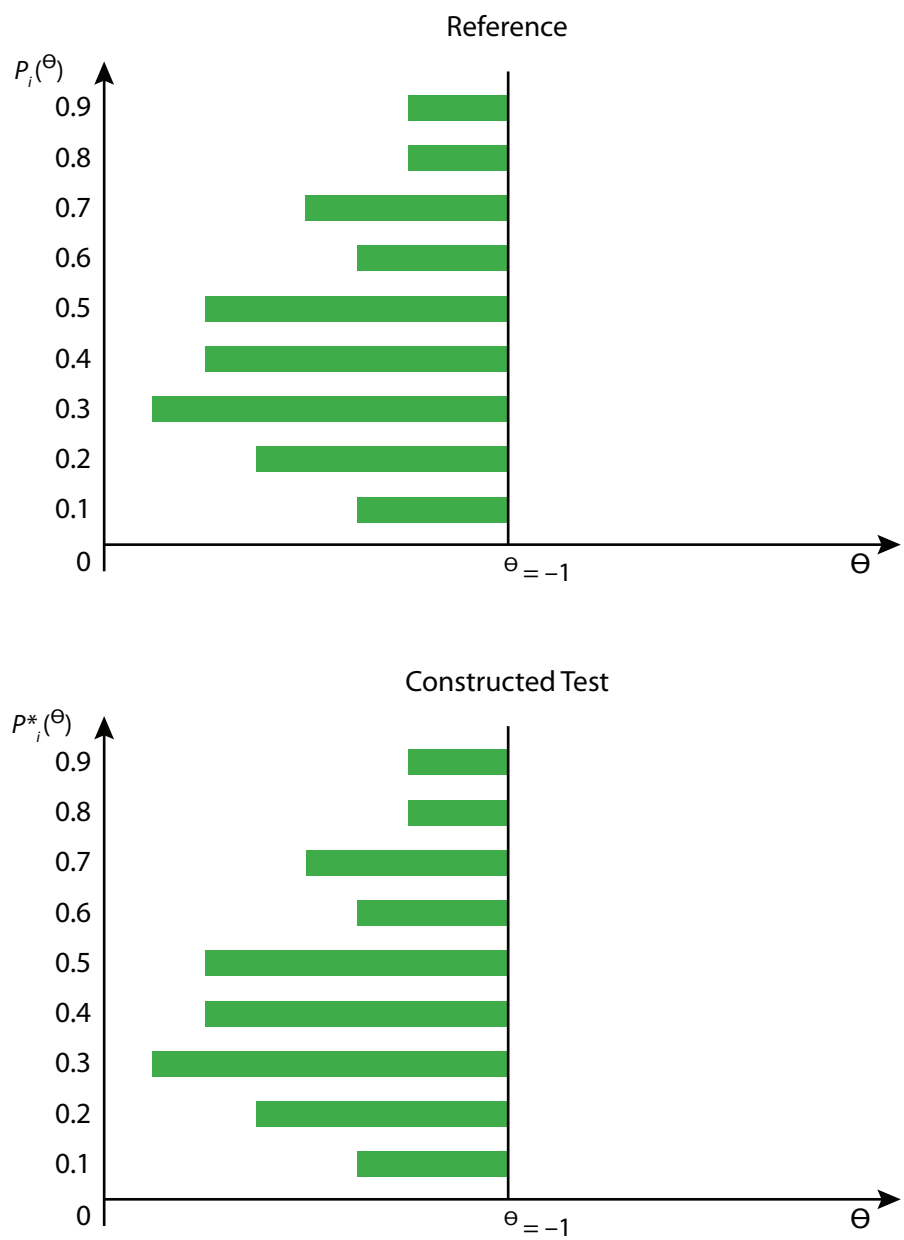
$f\ [P(\theta)]$ and $f\ [P^*(\theta)]$ are the density functions (as histograms or bars) at a particular $\theta$ level for the reference test and the constructed test, respectively, and $k$ denotes the number of bins formed for describing the histograms. The value of $OVR(\theta)$ lies between 1.0 (signifying complete overlap) and 0.0 (signifying no overlap). For example, two density functions share half of their density when $OVR(\theta)100\% = 50\%$. Figure 1 (next page) represents $OVR(\theta)100\% = 100\%$ for $\theta = -1$. If this result generates across the entire distribution of $\theta$, the constructed test can be considered to be strongly parallel (van der Linden and Luecht, 1998) to the reference test. Note that the lengths of the bars in Figure 1 represent $f\ [P(\theta)]$ and $f\ [P^*(\theta)]$ as stated in Equation 1. For an entire distribution of $\theta$, the expected value of $OVR(\theta)$, or $E_\theta\{OVR(\theta)\}$, can be considered. If it is assumed that the density of $\theta$ is $g(\theta)$, then

**Equation 2:**

$$E_\theta\{OVR(\theta)\} = \int_{-\infty}^{\infty} OVR(\theta) g(\theta) d\theta.$$

The value of $E_\theta\{OVR(\theta)\}100\%$ was computed by assuming that $g(\theta)$ is $N(0,1)$, where $E_\theta\{OVR(\theta)\}$ is the expected value of $OVR(\theta)$ for an entire distribution of $\theta$ used to signify the degree of statistical parallelism between any two tests, and also called Item Characteristic Curve Parallelism *(ICCP)* index in this study.

**Figure 1: Complete Overlap of the Conditional Distributions of $P(\theta)$ at $\theta = -1$ for the Reference Test and the Constructed Test**



*Smirnov Statistic T*

The Smirnov $T$ is defined as the maximal absolute vertical distance between two distribution functions (Conover, 1980). For this study, the Smirnov $T$ was used to describe the similarity in conditional distributions of $P(\theta)$ between a reference and a constructed test by measuring the maximal absolute vertical distance between the two distributions of $P(\theta)$. The measure of the maximal absolute vertical distance between the two distributions of $P(\theta)$ at a given $\theta_k$ was defined, in this study, as $MD(\theta_k) = Max[f\ \{P(\theta_k)\},\ f\ \{P^*(\theta_k)\}]$, where $f\ [P(\theta_k)]$ and $f\ [P^*(\theta_k)]$ are the density functions (as histograms) for the reference test and the constructed test, respectively, and $k$ denotes the number of bins formed for describing the histograms. The value of $MD(\theta)$ can range from 0 to 1. Smaller values of $MD(\theta)$ indicate greater similarity between two distributions. For an entire distribution of $\theta$, the expected values of $MD(\theta)$ or $E_\theta\{MD(\theta)\}$ can be considered. If it is assumed that the probability density function (i.e., pdf) of $\theta$ is $g(\theta)$, and $-\infty < \theta < \infty$, then

$$E_\theta\{MD(\theta)\} = \int_{-\infty}^{\infty} MD(\theta)g(\theta)d\theta.$$

$E_\theta\{MD(\theta)\}$ was used in this study to describe the degree of discrepancy (or similarity) between any two tests, assuming that $g(\theta)$ is N(0,1). Tests had greater *ICCP* were expected to yield smaller Smirnov $T$ values.

*Test Difficulty and Variability Indices*

The test difficulty indices included form difficulty or observed test mean, and the first central moment of $P(\theta)$ or test characteristic function. The test variability indices included form variability or observed test standard deviation, and the square root of second central moment of $P(\theta)$ or standard deviation of $P(\theta)$.

*Conditional Error Variance of Observed Test Score X (CEV of X)*

The *CEV* of $X$ is important in examining the conditional reliability of the constructed tests at each defined proficiency level. The *CEV* of $X$ at a single level of $\theta$ is defined as the sum of product of the conditional correct-response probability and the complement probability over all items in a test at that $\theta$ value, or

$$\sum_{i=1}^{n} P_i(\theta)\,[1-P_i(\theta)].$$

In conclusion, the criteria used to evaluate test parallelism included test overlap rate, content parallelism, and statistical parallelism. Statistical parallelism was assessed using a variety of indices. These indices included item characteristic curve parallelism *(ICCP)*, the Smirnov *T* statistic, the observed test score mean, and the observed test score standard deviation. Several additional indices conditional on the proficiency scales also were examined. These conditional indices included the first central moment of $P(\theta)$, the second central moment of $P(\theta)$, and the conditional error variance of the observed score.

# Results

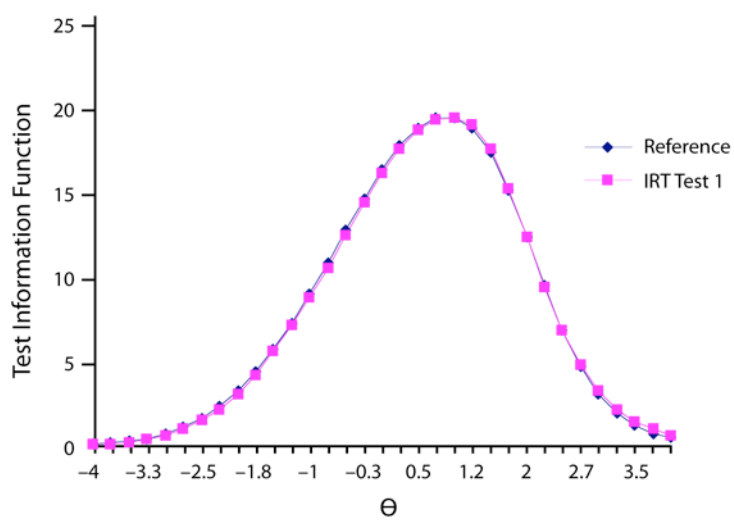
## Conformity to the Automated-Test-Assembly (ATA) Constraints

To ensure valid interpretation of the results, it is important to first consider the extent to which the test assembly approaches satisfied the imposed constraints. Psychometric constraints were met for the CTT and IRT-TTCF2P approaches but not for the IRT-TTIF approach. Table 5 (next page) shows the results for conformity to the psychometric constraints under each ATA approach. The term "Yes" in the tables indicates that all constraints were met, whereas "No" means that not all constraints were satisfied. However, Figure 2 (next page), displaying the information plots for the target test and the poorest matching test form generated with IRT-TTIF approach, showed that the information plots for the generated and target tests were very similar throughout most regions of the proficiency scale. The differences in test information between the target and generated tests were negligible and were not considered substantial enough to invalidate the IRT-TTIF approach and subsequent results for answering the research questions posed in this study. Content constraints were met for all ATA approaches. Therefore, all constructed test forms produced 100% content parallelism *(CP)*.

**Table 5: Results for Conformity to the Psychometric Constraints and the Percentages of the Unique Items in the Assembled Forms under the CTT, IRT-TTIF, and IRT-TTCF2P Methods**

| Form # | Met All Targets? | | | % Unique | | |
|---|---|---|---|---|---|---|
| | CTT | IRT-TTIF | IRT-TTCF2P | CTT | IRT-TTIF | IRT-TTCF2P |
| 1 | Yes | No | Yes | 90.00% | 81.67% | 86.67% |
| 2 | Yes | No | Yes | 86.67% | 83.33% | 98.33% |
| 3 | Yes | No | Yes | 83.33% | 81.67% | 88.33% |
| 4 | Yes | No | Yes | 83.33% | 78.33% | 95.00% |
| 5 | Yes | No | Yes | 91.67% | 78.33% | 88.33% |
| 6 | Yes | No | Yes | 90.00% | 80.00% | 91.67% |
| Average | NA | NA | NA | 87.50% | 80.56% | 91.39% |
| *SD* | NA | NA | NA | 3.62% | 2.02% | 4.52% |

**Figure 2: Test Information Functions for the Reference Test and IRT-TTIF Poorest Matching Test Form**



## Average Test Overlap Rate *(ATOR)*

After the three ATA methods were judged to conform acceptably to their constraints and produce 100% content parallelism *(CP)*, the methods were compared in terms of test overlap rate and statistical parallelism.

Table 6 shows the *ATOR* and *E*(*BOR*) results for each ATA approach. The results indicate that only the CTT and IRT-TTCF2P methods yielded acceptable test overlap rates. The *ATOR* for the CTT method was 0.172, which was greater than the *E*(*BOR*) by only 0.172 – 0.141 = 0.031 units when the statistical constraints (test difficulty and observed-score variability) were added. As for the IRT-TTCF2P method, the *ATOR* was 0.129, which was smaller than but pretty close to the *E*(*BOR*) by 0.141 – 0.129 = 0.012 units when the statistical constraints (TCF) were added. Therefore, no item exposure control would have been necessary using the CTT and IRT-TTCF2P approaches.

**Table 6: Average Results for Test Overlap Rate under the CTT, IRT-TTIF, and IRT-TTCF2P Methods (Deviation = *ATOR* – *E*(*BOR*))**

| Method | *E(BOR)* | *ATOR* | Deviation |
|---|---|---|---|
| CTT | 0.141 | 0.172 | 0.031 |
| IRT-TTIF | 0.141 | 0.579 | 0.438 |
| IRT-TTCF2P | 0.141 | 0.129 | –0.012 |

On the other hand, the *ATOR* for the IRT-TTIF method was 0.579, which was greater than the *E*(*BOR*) by 0.579 – 0.141 = 0.438 units when the statistical constraints (TIF) were added. Therefore, item exposure control would have been necessary using the IRT-TTIF method.

Greater *ATOR* signifies that the item pool may not have sufficient items to produce alternate test forms with no common items. If the number of test-assembly constraints get larger, the item pool may not support well the construction of unique alternate forms. Under those conditions, it was of interest to determine whether most of the items on the reference tests would be selected, and thereby to learn more about the performance of the WDM heuristic. Accordingly, the percentages of items that were unique in the assembled forms (i.e., did not appear on the reference form) were noted as a supplemental information (Table 5, previous page). As expected, the IRT-TTIF method had smaller percentages of unique items than the CTT and IRT-TTCF2P approaches since it had more test-assembly constraints.

## Statistical Parallelism

### Item Characteristic Curve Parallelism (*ICCP*) and Smirnov Statistic *T*

*ICCP* was defined as the expected percent of overlap between two conditional distributions of $P(\theta)$ over the entire distribution of $\theta$. This index represents a very strict test of statistical parallelism because achieving 100% parallelism would require one to one congruence of item characteristic functions across test forms. Table 7 shows the *ICCP* results for the six test forms created with each ATA method. In general, *ICCP* tended to be higher with the CTT and IRT-TTCF2P methods and lowest with the IRT-TTIF approach, but their differences were negligible. Nevertheless, the IRT-TTIF approach generated greater *ATOR* even though it yielded similar degrees of statistical parallelism to the other methods.

Table 7 also shows the Smirnov statistic *T* results for the six test forms created with each ATA method. Smirnov statistic *T* tended to be smaller with the CTT and IRT-TTCF2P methods and the greatest with the IRT-TTIF approach. The results for the Smirnov statistic *T* appeared to be in line with those for *ICCP* because tests that had greater *ICCP* were anticipated to yield smaller Smirnov *T* values and vice versa.

**Table 7: Item Characteristic Curve Parallelism *(ICCP)* and Smirnov Statistic *T* for Six Test Forms Assembled with the CTT, IRT-TTIF, and IRT-TTCF2P Methods**

| Form # | *ICCP* | | | Smirnov | | |
|---|---|---|---|---|---|---|
| | CTT | IRT-TTIF | IRT-TTCF2P | CTT | IRT-TTIF | IRT-TTCF2P |
| 1 | 48.84% | 46.54% | 49.85% | 0.116 | 0.199 | 0.086 |
| 2 | 50.63% | 47.19% | 46.74% | 0.111 | 0.194 | 0.111 |
| 3 | 52.98% | 49.58% | 50.73% | 0.090 | 0.178 | 0.144 |
| 4 | 50.67% | 48.56% | 46.67% | 0.110 | 0.191 | 0.131 |
| 5 | 46.36% | 49.92% | 47.67% | 0.128 | 0.206 | 0.120 |
| 6 | 49.20% | 47.48% | 48.77% | 0.100 | 0.198 | 0.167 |
| Average | 49.78% | 48.21% | 48.41% | 0.109 | 0.194 | 0.126 |
| *SD* | 2.22% | 1.36% | 1.67% | 0.013 | 0.010 | 0.028 |

### Form Difficulty – Observed Test Mean

Table 8 (next page) provides the observed test means for the assembled forms and their deviations from the reference test mean (i.e., 33.077) under the CTT, IRT-TTIF, and IRT-TTCF2P conditions. The average observed test mean over the six test forms for the CTT, IRT-TTIF, and IRT-TTCF2P methods were 33.158, 31.058, and 33.346, respectively. Additionally,

the corresponding average differences of the observed and reference test means over the six test forms were 0.081, –2.019, and 0.269. The CTT and IRT-TTCF2P test forms were not substantially different from the reference form in observed test mean. On average, the IRT-TTIF test forms had lower observed test means than the reference form. These results indicated that compared to the other methods, the IRT-TTIF method not only performed the worst in matching the reference form difficulty but also produced consistently more difficult tests (i.e., 1.654 to 2.212 total-score units below the target).

**Table 8: Observed Test Means and the Corresponding Deviations from the Reference Test under the CTT, IRT-TTIF, and IRT-TTCF2P Methods (Deviation = Assembled Test – Reference Test)**

| Form # | Expected Observed Test Score | | | Deviation (Reference = 33.077) | | |
|---|---|---|---|---|---|---|
| | CTT | IRT-TTIF | IRT-TTCF2P | CTT | IRT-TTIF | IRT-TTCF2P |
| 1 | 33.288 | 31.136 | 33.479 | 0.211 | –1.941 | 0.402 |
| 2 | 32.765 | 30.865 | 33.691 | –0.312 | –2.212 | 0.614 |
| 3 | 33.310 | 30.941 | 33.056 | 0.233 | –2.136 | –0.021 |
| 4 | 33.326 | 31.423 | 32.951 | 0.249 | –1.654 | –0.126 |
| 5 | 33.045 | 31.077 | 33.586 | –0.032 | –2.000 | 0.509 |
| 6 | 33.211 | 30.906 | 33.313 | 0.134 | –2.171 | 0.236 |
| Average | 33.158 | 31.058 | 33.346 | 0.081 | –2.019 | 0.269 |
| Avg. (abs) | 33.158 | 31.058 | 33.346 | 0.195 | 2.019 | 0.318 |
| *SD* | 0.218 | 0.207 | 0.295 | 0.218 | 0.207 | 0.295 |

Note: Avg. (abs) = average of the absolute values

### Form Variability – Observed Test Standard Deviation

Table 9 on the next page, shows the observed test standard deviations for the assembled forms and their deviations from the reference test standard deviation (i.e., 11.541) under the CTT, IRT-TTIF, and IRT-TTCF2P conditions. In general, the observed test standard deviations produced by each method were very similar to that of the reference test with differences ranging from 0.011 to 0.319. It is interesting to note that the CTT approach performed slightly better than the other approaches, but the differences among ATA approaches were generally small.

### The First Central Moment of $P(\theta)$ – Test Characteristic Function (TCF)

Figure 3 on the next page shows the test characteristic curves of the reference test form and one test form assembled with each of the three ATA approaches. Only one form constructed from each ATA method was

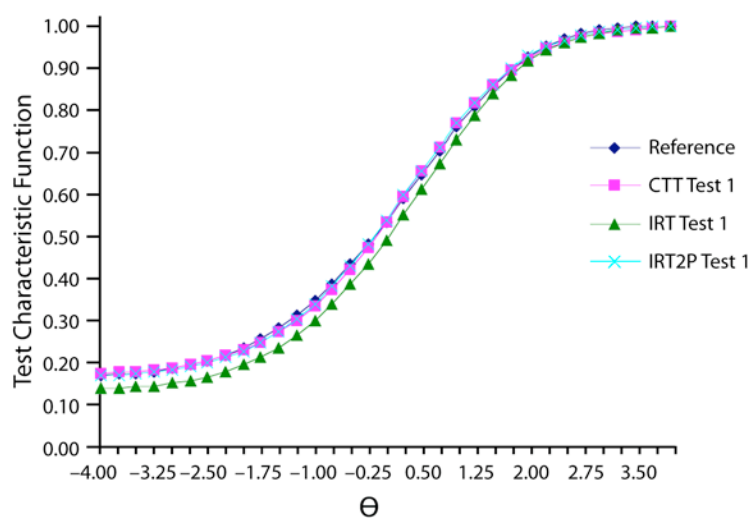
compared to the reference because the test characteristic curves were similar for the six constructed forms with each ATA method. In Figure 3, these test characteristic functions appeared to be similar, except that of the test form assembled using the IRT-TTIF method.

**Table 9: Observed Standard Deviations and the Corresponding Deviations from the Reference Test under the CTT, IRT-TTIF, and IRT-TTCF2P Methods (Deviation = Assembled Test – Reference Test)**

| Form # | Observed Standard Deviation | | | Deviation (Reference = 11.541) | | |
|---|---|---|---|---|---|---|
| | CTT | IRT-TTIF | IRT-TTCF2P | CTT | IRT-TTIF | IRT-TTCF2P |
| 1 | 11.45 | 11.725 | 11.254 | –0.091 | 0.184 | –0.287 |
| 2 | 11.439 | 11.779 | 11.526 | –0.102 | 0.238 | –0.015 |
| 3 | 11.83 | 11.755 | 11.395 | 0.289 | 0.214 | –0.146 |
| 4 | 11.6 | 11.759 | 11.415 | 0.059 | 0.218 | –0.126 |
| 5 | 11.552 | 11.86 | 11.646 | 0.011 | 0.319 | 0.105 |
| 6 | 11.66 | 11.798 | 11.318 | 0.119 | 0.257 | –0.223 |
| Average | 11.574 | 11.776 | 11.447 | 0.033 | 0.235 | –0.094 |
| Avg. (abs) | 11.574 | 11.776 | 11.447 | 0.110 | 0.235 | 0.136 |
| *SD* | 0.158 | 0.051 | 0.147 | 0.158 | 0.051 | 0.147 |

Note: Avg. (abs) = average of the absolute values

**Figure 3: Test Characteristic Functions for the Reference Test and Assembled Test Form #1 (IRT = IRT-TTIF, IRT2P = IRT-TTCF2P)**



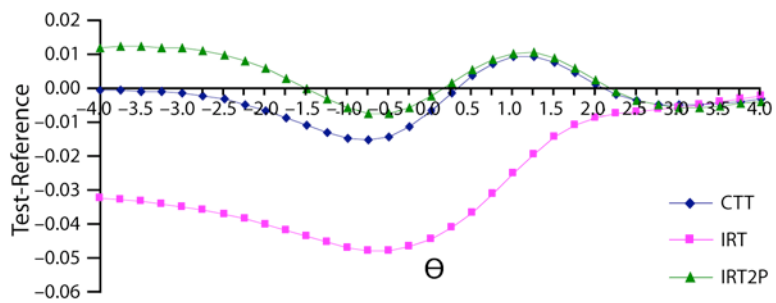
To provide an evaluation of differences in TCFs between the assembled and reference tests, the squared root sum of squared TCF deviations (between the reference and assembled tests) across all proficiency points was derived for the six test forms generated with each method (Table 10). This index provides an indicator of TCF deviation aggregated over the entire proficiency scale, and the smaller values represented greater conformity to the reference TCF. The results indicated that the CTT and IRT-TTCF2P approaches performed better than the IRT-TTIF approach in matching the reference TCF.

**Table 10: Aggregated Deviations between the Reference and Assembled Tests in the First Central Moment of *P*(θ), the Square Root of the Second Central moment of *P*(θ), and the Conditional Error Variance of Observed Score under the CTT, IRT-TTIF (IRT), and IRT-TTCF2P (IRT2P) Methods (Deviation = Assembled Test – Reference Test)**

| Form # | First Central Moment of *P*(θ) | | | Square Root of Second Central Moment of *P*(θ) | | | Conditional Error Variance of Observed Score | | |
|---|---|---|---|---|---|---|---|---|---|
| | CTT | IRT | IRT2P | CTT | IRT | IRT2P | CTT | IRT | IRT2P |
| 1 | 0.038 | 0.182 | 0.046 | 0.141 | 0.090 | 0.142 | 1.721 | 5.878 | 0.969 |
| 2 | 0.056 | 0.186 | 0.059 | 0.122 | 0.094 | 0.155 | 2.103 | 5.764 | 2.475 |
| 3 | 0.078 | 0.207 | 0.054 | 0.126 | 0.092 | 0.242 | 2.613 | 6.716 | 3.718 |
| 4 | 0.055 | 0.157 | 0.043 | 0.123 | 0.107 | 0.155 | 2.150 | 5.155 | 2.743 |
| 5 | 0.061 | 0.184 | 0.055 | 0.141 | 0.112 | 0.161 | 3.118 | 5.737 | 2.122 |
| 6 | 0.048 | 0.194 | 0.075 | 0.143 | 0.096 | 0.251 | 2.113 | 6.183 | 5.064 |
| Average | 0.058 | 0.183 | 0.052 | 0.131 | 0.099 | 0.171 | 2.341 | 5.850 | 2.405 |
| SD | 0.014 | 0.018 | 0.007 | 0.009 | 0.010 | 0.040 | 0.537 | 0.560 | 0.998 |

To reveal more detailed information about TCF congruence, the differences in TCFs between the reference and the assembled tests (assembled test – reference test) at selected proficiency points with each method were plotted in Figure 4. Each difference curve represented the average difference over the six test forms constructed with each ATA approach. In Figure 4, the horizontal line at a vertical axis value of 0 represented the reference test form. Compared to the reference form, the test forms constructed with the CTT and the IRT-TTIF2P methods appeared to have been more difficult for the examinees at lower ability (slightly below 0 on the θ scale) and easier for the examinees with higher ability (slightly above 0 on the θ scale). Figure 4 further reveals that the most noticeable TCF difference was found for the IRT-TTIF approach, and this method produced tests more difficult than the reference test over almost the entire proficiency scale. As a result, the IRT-TTIF approach performed the poorest among ATA approaches.

**Figure 4: Average Differences of Test Characteristic Functions between the Reference and Assembled Tests (IRT = IRT-TTIF, IRT2P = IRT-TTCF2P)**

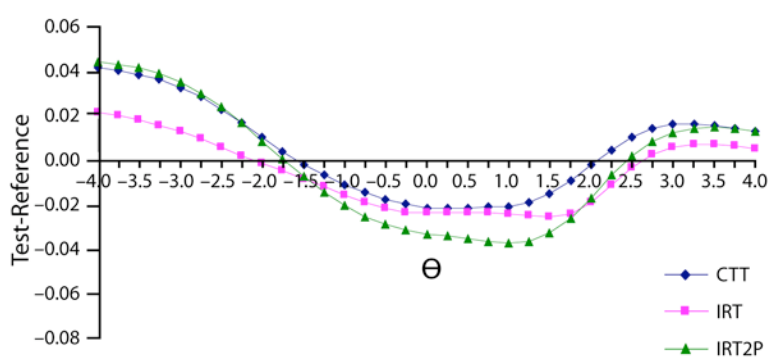


### The Second Central Moment of $P(\theta)$

To provide information about the congruence among ATA approaches, the aggregated deviations were also calculated for the standard deviation of $P(\theta)$ under each ATA method (Table 10). However, there was no compelling evidence that the CTT method did better than the IRT methods as the average deviation of the CTT method (0.131) was smaller than that of the IRT-TTCF2P method (0.171) but greater than that of the IRT-TTIF method (0.099).

To provide more detailed information about differences in congruence over a span of proficiency points, the conditional deviations (assembled test – reference test) are plotted in Figure 5. Again, each difference curve represented the average difference over the six test forms constructed with each ATA method, and the horizontal line at a vertical axis value of 0 represented the reference test form. The deviation plots for each ATA method had the same pattern. The standard deviation of $P(\theta)$ of each test form was smaller than that of the reference test form over the middle part of the ability range, but greater than that of the reference test form at both ends of the ability scale. Furthermore, the CTT approach only performed the best over the middle part of the ability range because its curve was the nearest to the horizontal line within that range, but it did not perform as well as the IRT approaches at the higher and lower ability levels. In general, similar to the results for aggregated deviations, the performance of the CTT approach could be regarded to be at least as good as the IRT approaches.

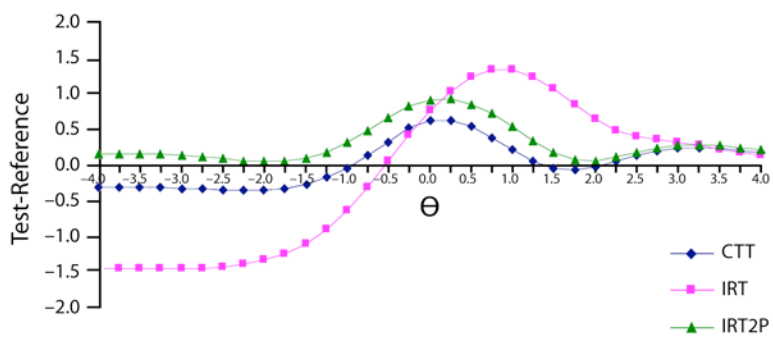
**Figure 5: Average Differences in Conditional Standard Deviation (or the Square Root of Second Central Moment) of $P(\theta)$ between the Reference and Assembled Tests (IRT = IRT-TTIF, IRT2P = IRT-TTCF2P)**



### The Conditional Error Variance of Observed Test Score *X* – *CEV* of *X*

Table 10 (page 29) and Figure 6 show the differences between the reference and generated tests in the *CEV* of *X* aggregated and conditional on the proficiency scale, respectively. Based on the information provided in Table 10, the CTT and IRT-TTCF2P approaches yielded, on the average, smaller aggregated deviations in the *CEV* of *X* (2.341 and 2.405, respectively) than the IRT-TTIF method (5.850). Similarly, in Figure 6, the difference curves generated from the CTT and IRT-TTCF2P approaches seemed to be closer to the horizontal line than the IRT-TTIF method over the entire span of proficiency levels. Accordingly, the CTT and IRT-TTCF2P approaches appeared to perform better than the IRT-TTIF approach in terms of conformity to the reference *CEV* of *X*.

**Figure 6: Average Differences in Conditional Error Variance of *X* between the Reference and Assembled Tests (IRT = IRT-TTIF, IRT2P = IRT-TTCF2P)**



# Discussions and Conclusions

## Discussion of Results

### Conformity to Constraints

Overall, the results of this study were in line with Stocking et al. (1993). The WDM heuristic produced tests with desired properties under realistic testing conditions. The content specifications (or constraints) were matched by all automated-test-assembly (ATA) approaches (i.e., CTT, IRT-TTIF, and IRT-TTCF2P approaches). The CTT and IRT-TTCF2P approaches also always satisfied its statistical constraints in that each assembled test form had a mean and standard deviation within the ± 0.5 boundaries. In contrast, the IRT-TTIF approach failed to meet the psychometric constraints in many instances. In these cases, the generated test information functions (TIFs) did not lie within the specified boundaries of the TTIF. However, the violations of the constraints were not considered serious enough to invalidate the IRT approaches. For example, the largest absolute difference between the generated TIF and TTIF (12.779 – 12.520 = 0.259) was negligible compared to the TIF of the corresponding constructed test form #1, 12.520, at $\theta = -0.5$. Failure to meet targets when using the IRT-TTIF approach may have been due in part to the arbitrary specification of boundaries for the targets. For example, the generated test properties (e.g., TIF) may fall out of bounds due to the narrow band specified for the targets. The decision of accepting a test missing such targets would be at the discretion of test developers and specialists based on their rationales and needs to be achieved. For this study, the conformity of the content and psychometric constraints was judged to be acceptable for all ATA approaches.

#### *Test Overlap Rate and Content Parallelism*

Test overlap rate is an important index to consider in ensuring the security of test items. Acceptable overlap rates were found for the CTT and IRT-TTCF2P methods. In evaluating these results, it should be emphasized that test overlap rates were simply observed as an outcome variable in the study. The high overlap rates for the tests assembled using the IRT-TTIF approach suggest that the overlap rate would need to be explicitly controlled to ensure adequate test security. However, if stringent test-overlap control is included as a test-assembly constraint, other content and psychometric constraints may need to be sacrificed to meet that constraint. In these cases, the value of $E(BOR)$ could be used as a benchmark for constraining test overlap rate that would be less likely to compromise test quality because this index ensures that content specifications will be met.

Content parallelism constraints were met 100% of the time for all ATA approaches. This result likely occurred because only four content constraints were imposed and the pool had a sufficient number of items to meet each constraint. However, such clear trends were not observed for statistical parallelism, which was examined using four global indices: Item Characteristic Curve parallelism *(ICCP)*, Smirnov Statistic *T*, observed test mean, and observed test standard deviation.

## Global Indices of Statistical Parallelism

The findings for the four global indices of statistical parallelism (Item Characteristic Curve parallelism *(ICCP)*, Smirnov Statistic *T*, observed test mean, and observed test standard deviation) varied by ATA approach. The tests produced by the CTT and IRT-TTCF2P approaches yielded questionable degrees of *ICCP*, but the *ATOR*s for those tests were acceptable. Those tests did not meet the strict standard of statistical parallelism for achieving one to one congruence of item characteristic function not only because the standard was difficult to reach but also because the tests were not assembled to meet that requirement initially. The IRT-TTIF approach yielded substantially greater (about five times greater) *ATOR* than the other two approaches even though all approaches produced similar findings for test parallelism. The greater *ATOR* for the IRT-TTIF approach may have been the result of the greater number of statistical constraints (i.e., 33) imposed for the IRT-TTIF approaches than for the other two methods (i.e., 2 for each).

Because the *ICCP* index may be an overly strict criterion for statistical parallelism, an alternative index (Smirnov statistic *T*) based on a less stringent criterion also was examined. The results for the Smirnov index showed that the CTT approach did even better than the IRT approaches in producing higher statistical parallelism. The *ICCP* and Smirnov indices yielded different results because they measure different statistical properties. *ICCP* reflects differences between distributions across the proficiency range, whereas the Smirnov statistic *T* only reflects the maximum deviations between distributions. Nevertheless, the outcomes for both indices lead to a conclusion that the CTT approach performed at least as well as the IRT approaches. It is particularly noteworthy that the CTT approach yielded higher parallelism than the IRT approaches even though the *ICCP* and Smirnov indices are IRT-based and would therefore seem to favor the IRT approaches.

The CTT approach also performed better than the IRT approaches in terms of the observed test mean and standard deviation criteria. This finding is not surprising given that the CTT approach constrained tests to have an observed test mean and standard deviation identical to the reference form.

Results for the observed test mean also yielded some interesting differences between the IRT approaches. In general, the tests assembled using the IRT-TTIF approach yielded greater differences with the corresponding reference tests in observed-score means than those constructed using the IRT-TTCF2P and CTT approaches. This trend occurred when all ATA approaches yielded questionable degrees of *ICCP*. These results likely occurred because the IRT-TTIF approach did not constrain the tests to have first-order (or first-moment) equity, whereas the CTT and the IRT-TTCF2P approaches did. Because the *ICCP* index was used to evaluate the equity of all moments, first-moment equivalence was not guaranteed. The same pattern of differences was shown for the observed-score standard deviation, but the differences among approaches were of smaller magnitude.

### Conditional Indices of Statistical Parallelism

To gain further possible insights into differences among ATA approaches, indices of central tendency (TCF), variability, and measurement error were examined conditionally on the proficiency scale. These conditional indices behaved differently with ATA approach. In general, the CTT and IRT-TTCF2P approaches produced similar patterns of TCF and/or TCF difference, but those patterns were different from that of the IRT-TTIF approach. The TCFs produced by the IRT-TTIF approach were substantially farther away from the reference TCF than those for the other two approaches. This result likely occurred because test difficulty (e.g., represented by TCF) was constrained by the CTT and IRT-TTCF2P approaches but not by the IRT-TTIF approach. Consistent with the results for global indices of statistical parallelism, the CTT approach performed as well as and in some cases better than the IRT approaches on conditional indices of test difficulty.

An important trend observed for TCFs was that the TCFs created by the IRT-TTIF approach were substantially more difficult than the reference test for examinees across most proficiency levels. Similar observations were made previously for the observed-score means in that the IRT-TTIF approach produced substantially lower means than the reference test. One possible reason that the IRT-TTIF approach created more difficult tests of medium difficulty is the correlation between IRT-based a and b parameters. That is, if the a/b correlations were greater in the pool than in the reference test, items of higher difficulty would be drawn from the pool and as a result tests assembled to match the reference test information function would be harder than the reference test. To evaluate this hypothesis, several medium difficulty reference tests with various a/b correlations were examined, but no clear trend was found between assembled test difficulty and the a/b correlations for the reference tests. Consequently, more research is needed to understand why this phenomenon occurred.

The important message from this finding for TCFs is that controlling test information does not guarantee equivalent levels of test difficulty even when the item pool contains sufficient items to support test assembly. This result seems reasonable given that a particular test information function could result from various combinations of item difficulties.

Results for the conditional variability of item difficulty and conditional error variance of the observed test scores were in line with those for the conditional index of central tendency. The CTT and IRT-TTCF2P approaches performed better than the IRT-TTIF approach. The IRT-TTIF approach yielded lower $P(\theta)$ variability curves and higher conditional error variance curves than those for the reference tests. These results make sense because the IRT-TTIF approach created more difficult tests with lower variability in item difficulties than those created by the other approaches.

## Summary of Discussion

Table 11, next page, summarizes the results and ratings for the CTT, IRT-TTCF2P, and IRT-TTIF methods under each test-parallelism evaluation criterion. A rating of 3 represents the best, 2 medium, and 1 the worst. The content specifications were matched for the CTT and IRT test-assembly methods. In general, the conformity of all constraints was evaluated to be acceptable for all ATA approaches. Since the CTT method obtained the most rating 3 and the IRT-TTIF method obtained rating 1 for all criteria except *SD* of $P(\theta)$, it appeared that the CTT performed the best whereas the IRT-TTIF performed the worst under the conditions specified in this study. To avoid a misleading answer to the primary research question, the comparison and conclusion of the findings for the two IRT methods were conducted first followed by those for the CTT versus IRT-TTCF2P methods. Special emphasis will be placed on noteworthy differences.

Results for *ICCP*, *ATOR*, the observed test mean, and TCF yielded some interesting differences between the IRT approaches. The tests produced by the IRT approaches yielded questionable degrees of *ICCP*, but the *ATOR*s were acceptable for the IRT-TTCF2P tests not for the IRT-TTIF tests. As stated previously, the greater *ATOR* for the IRT-TTIF approach may have been the result of its greater number of statistical constraints (i.e., 33). As for observed-score means, the tests assembled using the IRT-TTIF approach yielded greater differences with the corresponding reference tests than those constructed using the IRT-TTCF2P approach. Additionally, the TCFs produced by the IRT-TTIF approach were substantially farther away from the reference TCF than those for the IRT-TTCF2P approach. These results likely occurred because the IRT-TTIF approach did not constrain the tests to have first-order equity or equivalent test-difficulty (e.g., represented by TCF), whereas the IRT-TTCF2P approach did.

**Table 11: Summary of Results and Ratings for the CTT, IRT-TTCF2P (IRT2P), and IRT-TTIF (IRT) Methods under Each Test-Parallelism Evaluation Criterion**

| Criteria | | Results for Criteria | | | Rating | | |
|---|---|---|---|---|---|---|---|
| | | CTT | IRT2P | IRT | CTT | IRT2P | IRT |
| Conformity to Constraints | | yes | yes | acceptable | NA | NA | NA |
| Content Parallelism | | 100% | 100% | 100% | NA | NA | NA |
| Average Test Overlap Rate | | 0.172 | 0.129 | 0.579 | 2 | 3 | 1 |
| *ICCP* | | 49.78% | 48.41% | 48.21% | 3 | 2 | 1 |
| Smirnov Statistic *T* | | 0.109 | 0.126 | 0.194 | 3 | 2 | 1 |
| Average Deviation from Reference (absolute value) | Observed Test Mean | 0.195 | 0.318 | 2.019 | 3 | 2 | 1 |
| | Observed Test *SD* | 0.110 | 0.136 | 0.235 | 3 | 2 | 1 |
| | TCF | 0.058 | 0.052 | 0.183 | 2 | 3 | 1 |
| | *SD* of $P(\theta)$ | 0.131 | 0.171 | 0.099 | 2 | 1 | 3 |
| | *CEV* of *X* | 2.341 | 2.405 | 5.85 | 3 | 2 | 1 |

In many test programs and for many researchers, IRT methods are used exclusively to calibrate item response data and generate psychometric constraints. Under such situations, a target test information function and a target test characteristic curve at two values of $\theta$ might be specified together to create test forms not only having identical error variances (TIF) but also having equivalent true scores (TCF), which could fulfill the definition of test parallelism proposed by Lord and Novick (1968).

The ratings for almost all evaluation criteria tended to be greater with the CTT and lower with IRT-TTCF2P methods, but their differences in these indices were small or negligible. On the contrary, the IRT-TTCF2P method had a better rating than the CTT method for average test overlap rate and TCF but again the differences in these indices were very small. Furthermore, the finding for TCF is not surprising given that the IRT-TTCF2P approach constrained tests to have a TCF identical to the reference form. Taken as a whole, the results of *ATOR*, content parallelism, and various global and conditional indices of statistical parallelism lead to the conclusion that the CTT approach performed comparably with or better than the IRT-TTCF2P approach in automated assembly of parallel test forms. Therefore, under the situations which classical item statistics are the only data available, assembling parallel tests by constraining classical item statistics would be an appropriate way to solve the test assembly problem.

Taken together, the decision of using the CTT or IRT approaches to automate the assembly of parallel forms would be at the discretion of

test developers and specialists based on their rationales and needs to be achieved. However, when the IRT-TTIF method is selected for constructing tests, the item-exposure rate may need to be specified to reduce the item-reuse frequency for IRT-TTIF test forms, but the degree of parallelism may be worsened. Another way to decrease the item-reuse frequency for IRT-TTIF test forms is to relax the statistical constraints by broadening the bands surrounding the TTIF rather than specifying the item-exposure rate as a constraint. In other words, the information function of the assembled tests could be allowed to deviate from the TTIF more so that particular items would not be selected frequently to match the TTIF exactly.

## Conclusions

The purpose of this study was to investigate the degree of parallelism of test forms constructed with the WDM heuristic (Swanson & Stocking, 1993) using both classical and IRT approaches. This study was designed to answer a primary question: "Does the CTT approach perform as well as the IRT approach in the problem of parallel-test-form construction using the WDM heuristic?"

The general answers to the question are that the CTT approach performed as well as or better than the IRT approaches in assembling forms equivalent to the reference test, given that the item pool contained predominantly medium to slightly difficult items and the medium-difficulty reference test distribution was specified. That is, when the pool could supply adequate items for assembling parallel test forms, the CTT approach performed at least comparably with the IRT approaches in assembling parallel tests.

Based on the results of this study, the test forms assembled by each method could be considered to be pre-equated for the weaker definitions of parallel forms but not for the stricter definition of test parallelism. As stated in the introduction session, the goal of pre-equating is to derive equating transformations before test administration. If a particular definition of test parallelism (stricter or weaker) holds for the constructed forms, the goal of pre-equating can be considered to be achieved. McDonald (1999) proposed definitions for various degrees of test parallelism between test forms. For example, test forms are regarded as TIF-parallel if they have identical test information functions; ICC-parallel if they have identical item characteristic curves. Accordingly, test forms can be regarded as TCC-parallel if they have identical test characteristic curves, and regarded as test-mean-and-SD-parallel if they have identical observed test means and standard deviations. The three ATA approaches used in this study constrain the tests to have identical test means and standard deviations, TIFs, and TCFs, respectively. For this study, the conformity of the

constraints was judged to be satisfied or acceptable for all ATA approaches, and thereby the weaker definitions of parallel forms hold. It is interesting to note that the CTT and IRT-TTCF2P methods also yielded satisfactory degrees of test parallelism in terms of the psychometric criteria, except *ICCP*, used to evaluate test parallelism. On the other hand, the *ICCP* index in this study corresponds to McDonald's ICC-parallelism, which represents a strict definition of parallel forms. In this study, all ATA approaches produced tests with questionable degrees of *ICCP*, and thereby the stricter definition of test parallelism does not hold.

The CTT method appeared to work better in assembling forms equivalent to the reference test when the item pool contained predominantly medium to slightly difficult items and the medium-difficulty reference test distribution was specified. To investigate if this result can generalize to other conditions, in the future, a different study may be conducted with different types of item pools (e.g., item pool with higher or lower difficulty levels) and/or different sizes of item pools. Different types of statistical constraints may be specified together to produce test forms that match the requirements of particular testing programs.

## References

Ackerman, T. (1989, March). *An alternative methodology for creating parallel test Forms using the IRT information function.* Paper presented at the annual meeting of the National Council for Measurement in Education, San Francisco.

Armstrong, R.D., Jones, D.H., & Wang, Z. (1994). Automated parallel test construction using classical test theory. *Journal of Educational Statistics, 19*, 73–90.

Boekkooi-Timminga, E. (1990). The construction of parallel tests from IRT-based item banks. *Journal of Educational Statistics, 15*, 129–145.

Chen, S.-Y., 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.

Conover, W.J. (1980). *Practical nonparametric statistics*. New York, NY: John Wiley & Sons, Inc.

Gibson, W.M. & Weiner, J.A. (1998). Generating random parallel test forms using CTT in a computer-based environment. *Journal of Educational Measurement, 35*, 297–310.

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

Lord, F.M., & Novick M.R. (1968). *Statistical theories of mental test scores*. Reading, Massachusetts: Addison-Wesley.

Lord, F.M., & Wingersky, M.S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings". *Applied Psychological Measurement, 8*, 452–461.

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 Associates, Inc.

Spray, J.A., & Miller, T.R. (1992). *Performance of the Mantel-Haenszel statistic and the standardized difference in proportions correct when population ability distributions are incongruent*. (Research Report 92–1). Iowa City, Iowa: ACT, Inc.

Stocking, M.L., Swanson, L., & Pearlman, M. (1993). Application of an automated item selection method to real data. *Applied Psychological Measurement, 17*, 167–176.

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

van der Linden, W.J. (1987). Automated test construction using minimax programming. In W. J. van der Linden (Ed.), *IRT-based test construction* (pp.1–16). Enschede, The Netherlands: Department of Educational Measurement and Data Analysis, University of Twente.

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

van der Linden, W.J. (1998b). *Optimal assembly of tests with item sets*. (Research Report 98–12). Enschede, The Netherlands: Department of Educational Measurement and Data Analysis, University of Twente.

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., & Luecht, R.M. (1998). Observed-score equating as a test assembly problem. *Psychometrika, 63*, 401–418.

Zimowski, M., Muraki, E., Mislevy, R., & Bock, D. (2003). *IRT from SSI: BILOG-MG, MULTILOG, PARSCALE, and TESTFACT*. Chicago: Scientific Software international, Inc.

## Author Biographies

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