

Field Testing and Equating Designs for State Educational Assessments

Rob Kirkpatrick

Walter D. Way

Pearson

Paper presented at the annual meeting of the American Educational Research Association, New York,
NY, March 2008

Field Testing and Equating Designs for State Educational Assessments

Introduction

The educational accountability movement has spawned unprecedented numbers of new assessments. For example, the No Child Left Behind Act of 2002 (NCLB) required states to test students in grades 3 through 8 and at one grade in high school each year. Prior to NCLB, very few states had such comprehensive programs in place and the past several years have seen states and their contractors scrambling to implement the assessments needed to comply with the federal requirements. More recently, increased attention on college readiness has prompted many states to implement end-of-course examination programs (Achieve, Inc., 2007) to assess student performance in a variety of high school subjects. These new end-of-course programs have resulted in yet another wave of field testing and operational test implementation. Because each testing program operates under a set of policy, content, and development priorities that are unique to each state, the constraints that affect field testing and test equating may also be unique. As a result, measurement professionals routinely face difficult challenges when new statewide testing programs are started, such as:

- Customer requirements for immediate score reporting in the operational test, which can only be achieved through an item preequating design (Kolen & Brennan, 2004);
- Requirements in state law that every item on an operational form must be released to the public following the test administration;
- Customer requirements that a new test composed of multiple-choice and constructed response items must be initially field tested within 45-minute class periods;
- Customer requirements that tests be composed of multiple item formats, including innovative item formats, with items spanning broad domains;
- Passage-based tests that include a variety of extended response item formats, making an embedded field testing design infeasible.
- Customer requirements to field test and administered mixed-format tests in both paper-and-pencil and computer-based formats

There are numerous sources in the measurement literature that discuss and advise on test development and equating practice (c.f., Schmeiser & Welch, 2006; Holland & Dorans, 2006; Kolen & Brennan, 2004; Haladyna, 2004; Holland & Rubin, 1982). In general, these sources devote comparatively

little discussion to field test designs, the connection between field testing and equating, and associated technical issues. For example, Schmeiser and Welch (2006) describe two common approaches for field testing items, embedded field test administrations and special standalone studies. They advise that embedded field testing is generally preferred over standalone studies, citing test-taker motivation and test security as primary reasons. However, they stop short of providing specific examples of embedded or standalone field test designs. In their seminal book on test equating, scaling and linking, Kolen and Brennan (2004) discuss two equating designs that are sometimes utilized in state assessment programs, *common-item equating to a calibrated pool*, and *item preequating designs*. Their brief discussion of these designs is informative, and they conclude by acknowledging the many variations of designs that exist for equating using precalibrated item pools: “No attempt will be made here to enumerate all of these variations. However, context effects and dimensionality issues that arise with each variation need to be carefully considered when using item pools with operational testing programs” (pp. 207).

The purpose of this paper is to enumerate several of the variations related to field testing and equating that arise as testing programs are designed and implemented. There are two major sections. The first section discusses different field test designs and how these designs related to the use of field test item statistics (e.g., IRT item parameter estimates) in assembling and maintaining ongoing testing programs. Our discussion includes equating designs that may make use of item banks and field test item statistics. In the second part of the paper, we conduct a simulation to illustrate some of the issues and unintended consequences that may result from weak field testing and equating designs.

Categories of Field Test Designs

The primary purpose of field testing is to obtain information about item performance: distracter analyses, differential item functioning; and if subjective scoring is involved, rater-agreement information. A secondary purpose is to obtain statistics that can be used to assemble operational forms. Increasingly, field test data are used to obtain IRT item parameter estimates for establishing or replenishing a *calibrated item pool*. In certain situations, this purpose takes equal importance to that of evaluating item performance. There are two basic approaches to field testing: standalone and embedded (Schmeiser & Welch, 2006). In standalone field testing, examinees are recruited to take tests outside of the usual testing situation, and the test results are typically not used for instructional or accountability purposes. In many cases, all of the items appearing on standalone field tests are newly developed, and empirical data is being collected for the first time. In other cases, previously used items are placed on one or more standalone forms and used as linking items in order to place the IRT parameters on a common metric. We distinguish between the two forms of standalone field testing because they require different levels of commitment and expertise from the test developer. In our experience, inclusion of linking items requires additional

planning, experience with the target population, and psychometric and publishing expertise during the test construction phase. In embedded field testing, newly developed items are placed in an operational test, with several form variants being created. Figure 1 provides a graphical example of these three common approaches to field testing. Choice of which general design to use is based on the combined purposes and conditions of the project. In practice, there are many variations of these designs, ranging from simple to sophisticated.

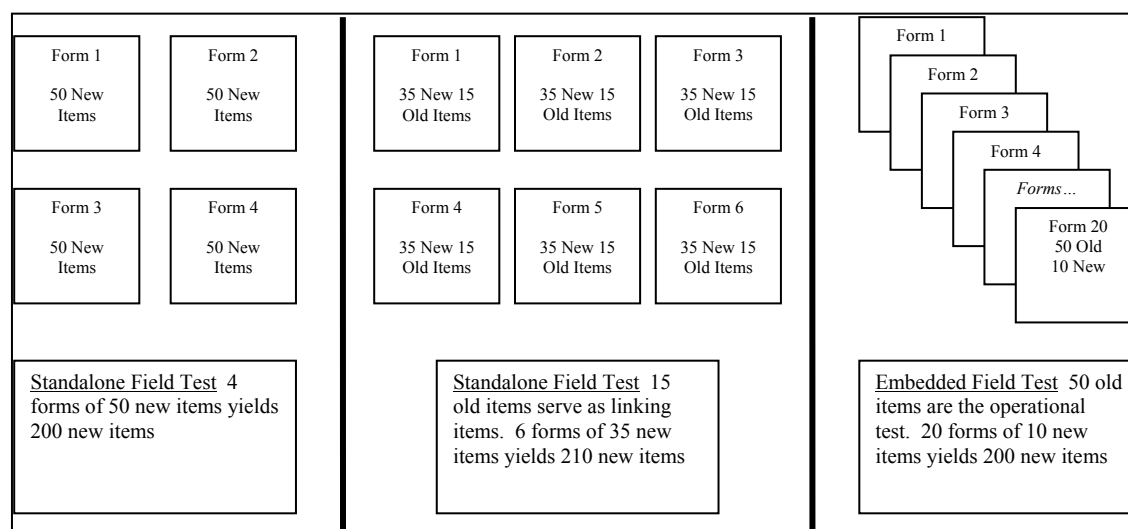


Figure 1. Three Approaches to Field Testing Approximately 200 New Items

The purposes and conditions of the testing program should drive the design decision. On the surface, embedded designs seem to have the most promise, but these designs also have significant constraints. Standalone designs provide the most flexibility as individual forms can be based on different blueprints, context-dependent sets are easily managed and can be paired for broader test design purposes, and publishing activities are simplified. Standalone designs also require special/additional testing sessions, are subject to motivation effects, and may not be the best approach for replenishing a *calibrated item pool*. On the other hand, embedded designs do not require special testing sessions and may reduce motivation effects since they are typically transparent to test takers and school officials. Embedded designs may not be available for new testing programs, may be difficult to justify for tests that are already long and subject to fatigue effects, pose significant problems for certain context-dependent test designs, and may be subject to item sequencing effects if used for replenishing a *calibrated item pool*. Table 1 lists some strengths and weaknesses of the standalone and embedded designs in common circumstances. As Table 1 illustrates, the relative strengths and weaknesses of these alternate designs depend upon the conditions under which they are being considered.

Table 1. Strengths and Weaknesses of General Field Test Designs

If the following conditions exist	Standalone		Embedded	
	Strengths	Weaknesses	Strengths	Weaknesses
Completely new program	<ul style="list-style-type: none"> • Flexibility • Fewest documents to publish. • Shorter test lengths possible 	<ul style="list-style-type: none"> • Requires randomly equivalent groups or linking strategy if IRT outcomes are important • Motivation effects 	<ul style="list-style-type: none"> • None 	<ul style="list-style-type: none"> • May not be possible
Replacing an existing program	<ul style="list-style-type: none"> • Flexibility • New items do not clue old items • New item development approaches do not expose new or old items • Shorter test lengths possible 	<ul style="list-style-type: none"> • Requires randomly equivalent groups or linking strategy if IRT outcomes are important • Additional test administration • Motivation effects 	<ul style="list-style-type: none"> • No additional administration • Motivation effects may be controlled if new items are not exposed • Item parameters easily placed on existing common metric 	<ul style="list-style-type: none"> • Lengthens test • Section exposure affects both old and new items • Sophisticated publishing activity • Position effects possible • Fatigue effects possible
Ongoing program: motivation known to be a problem	<ul style="list-style-type: none"> • Flexibility • New items do not clue operational items • New items development approaches do not expose new or old items • Shorter test lengths possible 	<ul style="list-style-type: none"> • May not be realistic 	<ul style="list-style-type: none"> • No additional administration • Motivation effects may be controlled if new items are not exposed • Item parameters easily placed on existing common metric 	<ul style="list-style-type: none"> • Lengthens test • Section exposure affects both old and new items • Sophisticated publishing activity • Position effects possible • Fatigue effects possible
Ongoing program: context-dependent set issues (e.g., context-dependent sets must be field tested together)	<ul style="list-style-type: none"> • Flexibility • New items do not clue operational items • New items development approaches do not expose new or old items • Shorter test lengths possible 	<ul style="list-style-type: none"> • Requires either randomly equivalent groups or linking strategy if IRT outcomes are important • Additional test administration • Motivation effects 	<ul style="list-style-type: none"> • No additional administration • Motivation effects may be controlled if new items are not exposed • Item parameters easily placed on existing common metric 	<ul style="list-style-type: none"> • May require unreasonable lengthening of the test

Initial Field Test Designs

Initial field test designs require carefully analyzing a number of factors. In initial field tests, a new testing program is being created and the item pool is typically under development. In cases where a new program is replacing an old one, an item pool may be available but is either not robust enough to meet the requirements of the new program, or the item statistics do not represent testing under the new specifications. The questions in Table 2 can be answered to identify the usability of item statistics from field testing. If the answer is “no” to any question in the table, the conditions of field testing may not reasonably approximate operational conditions, and the item statistics may provide limited information.

Table 2. Pre-Design Questionnaire for Initial Field Testing

#	Question	Yes	No
1	Will students have the same motivation in field testing as operational testing?		
2	Is the sample of students taking the field test expected to be randomly equivalent to the population that will take the operational test?		
3	Is the administration time the same for field and operational testing?		
4	If multiple forms are developed, are forms spiraled?		
5	Is the population of students familiar with the item types and testing mode being used?		

Issues related to representative sampling or sample size often impact initial field testing designs. Obtaining a representative sample of students can be impacted by constraints being placed on the project. Examples include:

- Participation for districts, schools, or students is voluntary
- Districts or schools cannot be sampled in consecutive years
- Forms must be assigned by district or school – spiraling not possible
- Sample sizes are restricted to levels below those desirable for IRT parameter estimation

Accounting for issues associated with the overall sample being non-representative of the target population is difficult with a new program. In such cases embedded linking designs may simply not be feasible, and spiraling of old forms assumes randomly equivalent groups that do not exist. If evidence suggests that the sample may be non-representative, the test developer should investigate the effects on item statistics before using the values for decision making purposes, or forego decision making purposes until piloting in a representative sample can be conducted.

In many cases, the sample is expected to be representative of the population, but form assignment procedures may lead to non-random groups. This may occur when item or form exposure controls are in place. IRT common-item linking designs are sometimes used in these cases (Kolen & Brennan, 2006). Figure 2 provides a relatively simple multiple-link solution using a chain to place three forms on the same scale. In this example, Form 2 is placed on the Form 1 scale using Linking Set A, which creates Link X. Linking Set B is placed on the Form 1 scale through Form 2, which creates Link Y. Finally, Form 3 is placed on the Form 1 scale using Linking Set B, which creates Link Z. Implementing designs such as that illustrated in Figure 2 requires high levels of dependency on the available items, adherence to IRT assumptions, and sophisticated production capabilities in form publishing, shipping, and possibly in examinee pre-identification.

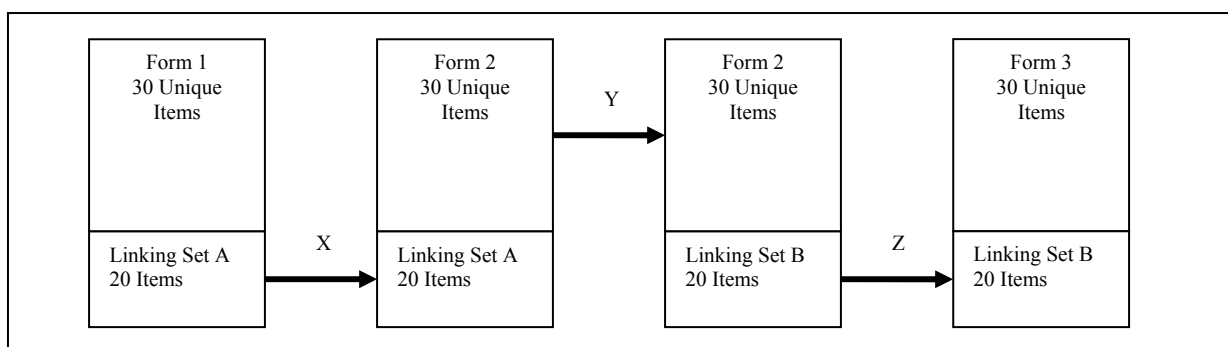


Figure 2. Example of Chain of Linking in Standalone Field Testing

If the conditions of field testing are similar to those of operational testing, then the IRT-based item statistics might be used to initiate a *calibrated item pool*, or decision-making such as setting performance standards. If at all possible, we recommend the field test results be compared with operational results before accepting the parameter values permanently into the calibrated pool, or publicizing the decisions made using those values.

Designs for Initial Test Construction

In many cases, a test developer is interested in field testing only a sufficient number of new items to meet very specific needs, e.g., to build a single operational form. Several factors need considered when designing the field test in support of initial test construction activities. Table 3 includes several questions test developers should consider before creating such a design.

Table 3. Questions for Initial Field Testing for Test Construction

#	Question	Additional Considerations
1	How many operational forms will be developed using this field test?	<ul style="list-style-type: none"> • Is ongoing field testing planned? • Are item statistics believed to have a shelf-life? • Does the program have scheduled item release? • Is item exposure a major consideration?
2	What is the expected item acceptance rate? E.g., what percentage of field test items is expected to demonstrate acceptable item statistics and be placed in the permanent item pool?	<ul style="list-style-type: none"> • Does one item type have a different rate than others? • Does one content area have a different rate than others?
3	How much time will be allowed for field test sessions?	<ul style="list-style-type: none"> • Is the administration time based on class sessions? • Is the allotted time significantly less than the planned operational form? • How long is the administration window in calendar days?
4	How many students will participate?	<ul style="list-style-type: none"> • Does participation rate affect the selection of IRT model, or vice versa? • Is there reason to expect the sample to be non-representative of the target population?
5	What is the unit of sampling and how is form assignment carried out?	<ul style="list-style-type: none"> • Is random spiraling of forms possible? • Is the IRT software capable of multi-group analysis?
6	What purposes are the data being used for?	<ul style="list-style-type: none"> • Scale construction? • Standard Setting? • Item evaluation? • Test construction? • Populating a <i>Calibrated Item Pool</i>?

Factors such as student motivation and administration time may have major implications for initial field test design. In our experience, students tend to be less motivated when the field test requires greater effort to complete. For example, when constructed-response items are administered in a stand-alone field test, it is not uncommon for very few students to score in the highest categories. This can be particularly problematic because field testing these human scored items also serves the purpose of establishing the scoring rules and rater training materials for future administrations. If the field test design is weak, some constructed response items might not be accepted into the pool because of insufficient numbers of high scores to build rater training sets.

Designs for Embedded Field testing

Our discussion of embedded field testing designs cover a variety of purposes, including new item tryout, external linking sets (e.g, equating anchors, vertical scaling), matrix sampling of extended content domains, and implementing research studies. There are two primary advantages of embedding: field testing takes place in the same administration as operational testing, and motivation effects are reduced or eliminated. These advantages come at some cost: the operational administration must be lengthened, and test assembly activities require expertise and test publishing activities are much more sophisticated. This field test design also has three potential weaknesses: item position and context effects, the potential introduction of dimensionality effects that are not consistent with the operational test, and the need to preserve the confidentiality of field test positions.

In practice, embedding has two common implementations: 1) embedding in an intact section and, and 2) embedding items throughout the test. The option to chose is based on the particular needs of the program and the publishing technology available. Table 4 lists some strengths and weaknesses associated with common approaches to embedding. We are compelled to note that there is a very real and meaningful trade-off between meeting the psychometric goals of embedding and costs associated with paper-and-pencil test publishing activities. Any embedding increases costs in test publishing, and sophisticated embedding may increase costs associated with scoring. Challenges to solve include managing clueing between field items and operational items, white space in form publishing, printer collation issues, and what to do with students who do not indicate their form on the answer document. In the future, these challenges will become much easier to address as testing programs move their assessments online. Online testing makes it possible to implement and monitor very sophisticated algorithms for administering field test items within operational tests.

Appending embedded sections at the end of a test book requires less effort and exposes the program to the fewest challenges related to the burden of field testing and its potential impact on student performance. However, it provides the least benefits from a psychometric perspective. Because of the potential for item position effects, section embedding at the end of a test may not be reasonable for *calibrated item pools*. We use simulations to illustrate this problem later in this paper.

Table 4. Strengths and Weaknesses of Common Embedding Strategies

Embedding Design	Strengths	Weaknesses
Items in a section at the end of the test	<ul style="list-style-type: none"> • Fatigue effects may be isolated in field items • Clueing less likely to be caused by field items than other options • The same sections can be appended to multiple operational forms • Field sections can vary in length, both in terms of number of items and in physical space • Less impact on publishing activities than other options • Quality control more easily managed 	<ul style="list-style-type: none"> • Maximum potential for item position effects • Difficult to leverage for external linking purposes • Maintaining security of sections may be challenging • Under some circumstances, may create local item dependence • May not be reasonable for <i>calibrated item pools</i>
Items in a section in the middle of the test	<ul style="list-style-type: none"> • Minimizes item position effects for context-dependent test designs • Less publishing expertise than embedding individual items 	<ul style="list-style-type: none"> • Fatigue effects may be isolated in operational items • Field items may clue operational items • If field sections vary in physical length, adds to publishing complexity • Under some circumstances, may create local item dependence
Items divided into multiple sections throughout the test	<ul style="list-style-type: none"> • Minimizes item position effects for discrete items or small context-dependent sets • Fatigue effects spread across all items (could also be seen as a weakness) • Sections can be used for other purposes • Field sections more difficult to identify 	<ul style="list-style-type: none"> • High degree of publishing sophistication • More quality control required • More potential for clueing
Items placed throughout the test as individual items	<ul style="list-style-type: none"> • Potential to completely control for item position effects 	<ul style="list-style-type: none"> • Highest degree of publishing sophistication • Most quality control required • Most potential for clueing • May not be reasonable for tests composed of context-dependent sets

To best avoid item position effects, items should be placed relatively close to their field position when they are used operationally. In tests with broad domains, it is important to consider the impact of embedding on item statistics. Overweighting one content area may create dimensionality inconsistencies between different field-test forms, and if forms are not randomly spiraled, may result in bias in item means. On the other hand, if items in the embedded section seem completely unrelated to one another, it

may result in section exposure to observant test takers. Periodically changing the location of embedded sections is recommended to manage exposure (Schmeiser & Welsh, 2006). If embedded items look different in any way from the operational items, section exposure may be unavoidable.

Designs for Ongoing Standalone Field Testing

There are practical reasons for ongoing field testing to use a standalone design, both from a test design and a practical perspective. Table 5 lists issues a test developer should consider to determine which approach would be more favorable for the program. The most salient issues are related to the operational test design, and whether or not item statistics must be used directly.

Table 5. Issues to Consider in Ongoing Field Testing: Embedded and Standalone Compared

#	Issue	Embedded	Standalone
1	Motivation effects must be managed	Preferred	Difficult
2	Frequent number of test administrations	Preferred	Difficult
3	<i>Calibrated Item Pools</i> approach to equating/preequating is used	Preferred	May not be reasonable
4	Representative samples cannot be acquired	May be better	Limited use of statistics
5	Operational test is considered long	Difficult	May be better
6	The operational test is timed	May be difficult	May be preferred
7	Operational test design includes theme or paired contexts	May not be possible	Preferred
8	Writing prompt or essay that is expected to take 30 or more minutes for students to complete	May not be reasonable	Preferred
9	Field sections are likely to be exposed	Benefits diminished	No preference

Although the advantages of embedding are compelling, there are operational test design features that may make it difficult or even unreasonable to use this design. Any test design that would require embedded sections to be prohibitively long, or add significant testing time to the operational administration is probably not a good candidate for embedding. This may be the case when programs utilize extended constructed-response items in their operational assessment or utilize overarching theme as a context of reference for most or all the items in the test.

Sampling typically requires more attention and planning in standalone field testing. Many states do not have authority to require districts or schools to participate. Even when authority is granted, there may be reasons to allow constituencies to opt-out of an administration, e.g., district is participating in the National Assessment of Educational Progress that year. When an operational test design is better suited to standalone field testing, the test directions or proctor instructions may be specific to the field test form.

As a result, sampling may need to take place at the classroom, school, or district level. This approach may also be preferred if item and test security are concerns. For practical reasons, all students at a location may be required to participate in the field test.

In ongoing field testing efforts involving a standalone design, a strategy is needed to link IRT parameters to the desired metric. The common item or random groups procedures defined in Kolen & Brennan (2004) are typical choices. If common item procedures are used, the common item sets must have IRT parameter estimates on the appropriate metric. Old operational forms spiraled in the field test administration can also be used. Figure 3 provides two examples of how this might be accomplished. Using old operational forms has the advantage of keeping the publishing activity simple. The test developer needs to be aware that these linking strategies are contaminated by confounding effects due to different standardization conditions, e.g., different motivation, different time of year, etc. If the conditions of the field test administration are dramatically different from that of the operational administration, issues may arise in the linking outcomes over time.

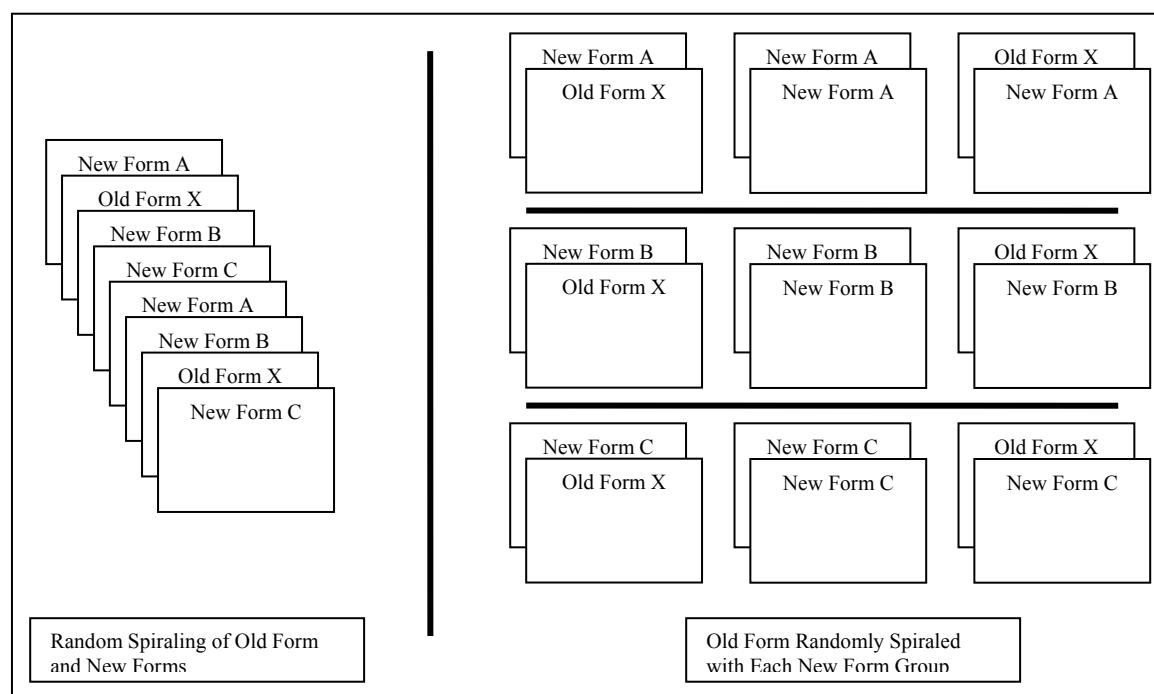


Figure 3. Two Examples of Spiraling an Old Form in a Field test Distribution

It may be possible to place the field-test item parameters on the operational metric using a common persons design rather than a common item or old form spiraling design. In the common persons design, standalone field tests are administered under as similar conditions as possible to the operational test, and within as close proximity in time as reasonable. After the operational test has been scaled, item parameters estimated and placed on the formal metric, the responses of students taking both the field test

and the operational test can be leveraged to link the field-test item parameters to the operational metric. Common person procedures assume that student motivation in both test administrations are identical. This assumption is likely to be violated in practice, especially if the operational test is associated with student-level stakes. Without using common items or old forms, it is not possible to know if the assumption of equal motivation was violated until after the new items are administered in an operational situation. Because of the motivation issue, we find the procedure risky, and recommend the use of other approaches if at all possible. If this procedure is the only option, the resulting item statistics probably should not be used in scaling or equating the operational test.

When standalone field testing is used for replenishing existing item pools in a high-stakes environment, test developers should attempt to replicate the operational conditions as closely as possible. For example, it may be possible to increase motivation in a standalone field test by offering student incentives for participation, although there are likely to be policy challenges with doing so.

Equating for *Calibrated Item Pools*

When operating a *calibrated item pool*, item parameter estimation becomes a primary purpose of field testing. Scale transformation procedures described in Kolen & Brennan (2004) are typically used to place the field-test item parameter estimates on the operational latent trait scale. Item parameters estimated under differential standardization conditions are most likely not on the same scale and should not be commingled. As such, the issues discussed above are all pertinent, and increase in importance. The task of equating to a *calibrated item pool* depends on whether the pool is newly created or ongoing. For newly created pools, the latent trait scale is sometimes established after the first operational administration rather than the field test administration, especially in cases where field test motivation effects are expected or observed. To place the item parameters from field test on the operational scale the items appearing on the first operational test can be use as links and a scale transformation applied. If this activity is needed, we recommend that item parameters obtained from field test be replaced with operational values once items are used on an operational test. A graphical example of this process is provided in Figure 4.

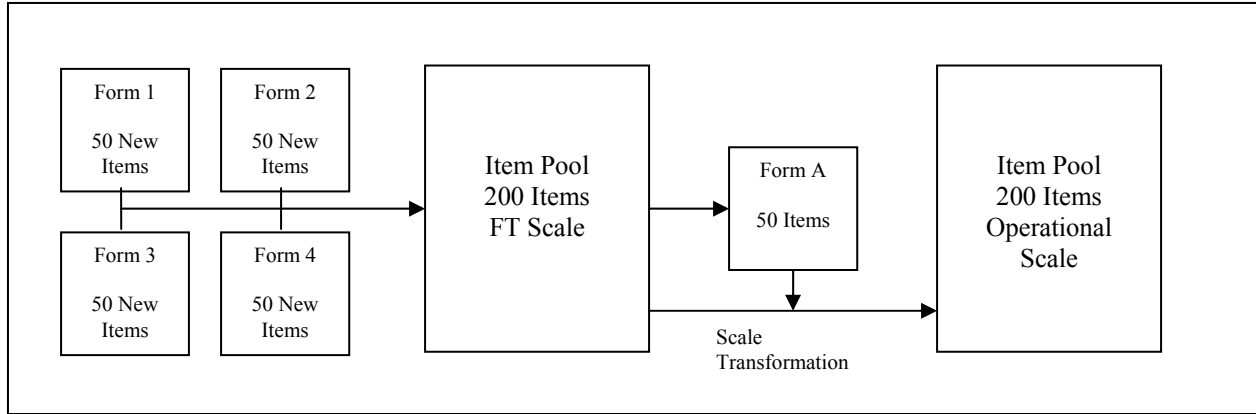


Figure 4. Initial Scaling for Calibrated Item Pool – Form A is First Operational Administration

In maintaining a *calibrated item pool* equating is conducted to place the item parameters of newly field-tested items on the operational latent trait scale. These item parameters are then used for test construction and score reporting. We note that a major benefit of using a *calibrated item pool* is the ability to create pre-equated test forms and use pre-equating for score reporting purposes. Equating strategies that apply post-equating for operational forms are common, however, they do not make full use of the *calibrated item pool* approach.

Assuming the scale is established in Year 1, and the *calibrated item pool* is on the operational scale, new test forms may be created and pre-equated using item parameters from the pool. Provided the test form is built according to content and statistical specifications, and all items are from the *calibrated item pool* the scores are considered equated. Field test items or sections are added to the test according to the defined design. The field-test item parameters are placed on the operational scale using the operational items as links (anchors) and a typical scale transformation procedure.

In post-equating, new operational data is obtained for items selected from the *calibrated item pool*. Item parameters are estimated for the operational data, and operational items are post-equated using the pool (old) and current (new) item parameters and a scale transformation procedure. If new field test items were administered with the operational items, this transformation can be applied to their calibration results as well.

Over time, both approaches result in a chain of equatings that lead away from the first year in which the scale was established. Since equating errors are additive, the further away from the base year, the more scale drift may occur. As such, it is important to include procedures that evaluate and manage scale drift. One procedure is to use a dual-linking strategy and leverage the field test positions for external common-item linking. Linking sets can come from Year 1 administration(s), or other old forms

to link as close to the base year as possible. The external link can be used as in the operational equating and/or to evaluate drift. One strategy that may be followed to monitor scale drift is to create several external links and establishing their item parameter values early in the program history, then rotate these links periodically in order to manage exposure. Conceptually, the difference between the scale transformation solutions using the operational items as links and the external set as links represents scale drift. Careful creation of the different links would be required to draw conclusions. If section exposure is a problem, more sophisticated techniques will be required.

Illustrative Simulation

Initial Data

To this point we have discussed several designs and strategies for field testing. Some design decisions have important psychometric impacts to consider before choosing the design. We turn now to an illustration of one design consequence that test developers should consider when using a *calibrated item pools* approach: item position effects in embedded designs. To illustrate these effects over multiple years of test equating, simulations were conducted using a set of 181 field test items field tested in a single year from a real large-scale testing program. The testing program measured a diversified mathematics domain and was in place for several years. Use of Rasch item difficulties from real field test items was intended to provide more realistic outcomes. For our simulations, the pool was considered to be typical of the field test pool that would be administered each year, and used as the basis for simulating effects over time. The mean, standard deviation, minimum, and maximum of the RIDs for the mathematics items were -.0005, 1.0000, -2.489, and 3.371, respectively. These items were originally spiraled in sets of 10 in different operational test form versions containing 65 items.¹ For the purposes of the simulation, we developed test assembly specifications typical of those used with the operational test. Specifically, we grouped the item difficulties of the pool into 14 different intervals of width .30 theta. A set of 50 items that was proportionally representative of the complete pool was randomly selected and treated as a prototypical operational form. Call this the Year 0 form. We then equated future forms to the Year 0 form. Rasch true abilities were established for the simulations using the item parameters for this form and true score equating procedures, specifically by finding the ability level that corresponded to each possible number right true score on the form. In addition, quadrature weights were generated for each true ability level based on the weights from the actual operational test.

Simulation Procedures

¹ Some items were dropped after review of the results, thus leading to 181 items.

The approach for our illustrative simulations was based on a study recently conducted by Meyers, Miller, and Way (2006; in press). Specifically, we used the regression-based equation developed in that research to predict how Rasch item difficulty changes as a function of change in the position of the item from one test administration to the next. We expanded on their simulations in two ways. First, we simulated equatings over three administrations to investigate how the effect can compound over time. Second, we simulated changes in item difficulty assuming two designs for embedding field test items, one where items are field tested in the middle of the test and another where field test items are appended at the end of the test.

To conduct the simulations, the following steps were completed with 100 iterations:

1. Sample, with replacement, a pool of 181 items from the original pool of 181 items. Treat the item parameter estimates as true for the first year of equating. Call this Year 1.
2. Create a set of “field test” item parameter estimates by “perturbing” (i.e. adding an error component to) the true parameters, with error components based on SE_b assuming a particular distribution of ability (θ) and sample size.

First, for the Rasch item difficulty (RID) of each field test item i , $i=1, \dots, 281$, compute its information across the 51 ability quadrature points as follows :

$$I(b_i) = \sum_{j=0}^{51} P_{ij}(1 - P_{ij})\pi_{\theta_j} N_{FT}$$

where

$$P_{ij} = [1 + e^{b_i - \theta_j}]^{-1}$$

b_i = "true" RID for item i

π_{θ_j} = assumed proportion of students with ability θ_j

N_{FT} = assumed number of students field tested

The standard error for each b_i is then computed as follows:

$$SE_{b_i} = 1 / \sqrt{I(b_i)}.$$

Finally, the perturbed "field test" item parameter estimate for each item is

$$\hat{b}_i = b_i + z_{random}(SE_{b_i})$$

where z_{random} = random variate from $N(0,1)$ dist

3. Construct a test by selecting at random 50 items from the new pool of 181 field test items across the difficulty strata that were established when selecting the prototypical operational form (Year 0). Call this Year 1. A small random noise value was temporarily added to the item difficulties, and then the items were sorted by this temporary difficulty. In this manner the selected items were ordered generally by item difficulty, with easy items coming first. The temporary difficulties were then discarded. Final operational item positions were assigned based on two approaches to inserting the field test items: middle or end. When the field test positions were in the middle, the operational positions were 1-25 and 36-60. Then the field test positions were at the end, the operational positions were 1-50.
4. For the items selected in step 3, create a set of "final-form" item parameter estimates by perturbing the true parameter components based on SE_b as in step 2, assuming a sample size that was five times the corresponding field test sample size. In addition, a fixed component consistent with Meyers, Miller, and Way (in press) was added to the true item difficulty that was calculated by entering the difference between each item's position field tested and final form item positions.

For the RID of each field test item i , $i=1, \dots, 181$, compute its information as follows :

$$I(b_i) = \sum_{j=0}^{51} P_{ij}(1 - P_{ij})\pi_{\theta_j} N_{FF}$$

where

$$P_{ij} = [1 + e^{\theta_j - b_i}]^{-1}$$

b_i = "true" RID for item i

π_{θ_j} = assumed proportion of students with ability θ_j

N_{FF} = assumed number of students tested with "Final Form"

The standard error for each b_i is then:

$$SE_{b_i} = 1 / \sqrt{I(b_i)}$$

The perturbed "final form" item parameter estimate for each item is:

$$\hat{b}_i = b_i + (0.00329\Delta + 0.00002173\Delta^2 + 0.00000677\Delta^3) + z_{\text{random}}(SE_{b_i})$$

where

Δ = final form item position - field test item position

z_{random} = random variate from $N(0,1)$ distribution.

5. Establish a “true effect” for examinee performance by modifying the true item parameters from step 1 using the prediction equation from Meyers, Miller & Way (in press):

$$badj_i = b_i + (0.00329\Delta + 0.00002173\Delta^2 + 0.00000677\Delta^3)$$

6. Link the final form item parameter estimates to the field test estimates by calculating the mean difference between the two sets of estimates. As part of the linking, an iterative “stability check” was utilized, through which items with differences between field test and final form difficulty estimates greater than 0.3 were eliminated from the linking item set. This step imitated operational procedures used in many testing programs, which are commonly employed in IRT test equating settings.
7. Generate a test characteristic curve (TCC) based on the final linked item parameter estimates and compare this to the true TCC based on the true item parameters.
8. Randomly draw a new set of 181 field test items, with replacement, from the original pool of 181 items. Perturb the item parameters following the procedures indicated in step 2. These represent the new field test items administered with the Year 1 test, and the basis for constructing the Year 2 test. Apply the equating constant from step 5 to these item parameter estimates.
9. Create and equate test forms representing Years 2 and 3 by following steps 3 through 8 two consecutive times.

Data Analyses

The outcomes of interest over the 100 replications included the true minus estimated TCC differences (evaluated over fixed theta levels), the differences between the mean of the true and final estimated item difficulties (i.e. the equating constant), and the number of items removed in the iterative stability checks across the 100 replications and 3 years of equating. These summaries provided information about the properties of the tests constructed. In addition, differences between true TCCs and the TCCs calculated in step 5 were calculated to document the “true effect” on performance due to changes in item position. To vary the values of SE_b for the simulations, two variations of field test and final sample sizes were used: 500 and 2,500, 2,000 and 10,000. For each sample size combination, two sets of 100 replications simulated the systematic error due to the calculated changes in item position (middle and end) and a third set of 100 replications did not incorporate any systematic error. In the third set of 100 replications, step 4 as described above included the $z_{\text{random}}(SE_{bi})$ component but did not include

the component based on the function of item position change. This resulted in 6 sets of replicated simulations in total.

Simulation Results

One outcome of perturbing the item difficulties was that, by chance, a randomly selected pool of items might not meet the specification for selecting operational forms. This outcome was realized a small number of times, and is partly responsible for the number of forms built per 100 iterations as reported in Table 6. The effects of item position change were generated using a regression equation derived by Meyers, Miller and Way (in press) from a different real testing situation. The regression equation may not have been completely applicable for this illustration. For example, if field test items were assumed to be placed in positions 51-60, the furthest an item could move from the field test to the operational test in our simulations was -59 positions (60 to 1). However, with the test analyzed in Meyers et al. (in press), few items moved more than 40 positions. Thus, we chose to truncate effects to the magnitude resulting from a change of 45 positions.² Using the regression equation, the maximum predicted change in item difficulty for such items is -.701. If the field test items are placed in positions 26-35, the maximum predicted item position change effect is 34 positions (25 to 60), which corresponds to a predicted difficulty change of .403. The impacts on equating for field testing in the middle versus the end were expected to be different. For items field tested at the end of the test, all items move to earlier positions in the operational test and are predicted to be easier than when field tested. However, for items field tested in the middle of the test, the expected effects due to change in position were less clear.

Table 6 presents a summary of the simulation results across the different sample sizes (500/2500 and 2000/10000, respectively), and item position change conditions (no change, middle, and end). The columns of the table present the means and standard deviations across iterations of the numbers of items remaining after the post-equating stability check, the equating constants, the average difference between true and estimated TCCs, and the number of iterations (out of 100) for which a successful test assembly was completed. In some iterations, the process of sampling items resulted in insufficient items within a difficulty strata to build a test. This outcome was more prevalent in the field test at the end condition, as will be discussed below.

As expected, no effects were found for the no item position change simulation condition. For this condition, the equating constants were zero on average, and the standard deviations of the equating constants across replications were very small. In addition, very few items were removed through the iterative stability check and the average differences between true and estimated TCCs was near zero

² Using the regression formula to predict item position change for -59 positions would have resulted in change in item difficulty of -1.509, which seemed more extreme of an effect than would be reasonable to expect in practice.

across each of the three years of equating. For the field test in the middle condition, small negative average equating constants were found. For example, in the 500/2500 condition, these average constants were -.009, -.020, and -.031, for years 1 through 3 respectively. These constants reflect the compound effects of item position change over time. Average differences between true and estimated TCCs in this condition also changed across years, moving from 0.072 to -0.080 in the 500/2500 condition. Results for the 2000/10000 condition followed trends similar to the 500/2500 condition.

Table 6. Summary of Simulation Results

Condition	Field Test		Items Remaining		Equating Constant		True-Estimated TCCs		Forms Built
	Position/Effects	Year	M(SD)	Min,Max	M(SD)	Min,Max	M(SD)	Min,Max	
500/2500	None	1	49.37(0.71)	47,50	0.00(0.017)	-0.04,0.03	0.002(0.001)	-0.001,0.003	99
		2	50.00(0.00)	50,50	0.00(0.018)	-0.045,0.036	-0.002(0.007)	-0.010,0.009	100
		3	49.98(0.20)	48,50	0.00(0.019)	-0.049,0.042	0.002(0.004)	-0.002,0.009	99
	Middle	1	45.61(1.71)	41,49	-0.009(0.024)	-0.076,0.048	0.072(0.289)	-0.326,0.483	100
		2	48.01(1.04)	45,50	-0.020(0.027)	-0.093,0.039	-0.012(0.293)	-0.413,0.412	100
		3	47.52(1.49)	43,50	-0.031(0.031)	-0.113,0.034	-0.080(0.297)	-0.481,0.355	99
	End	1	33.11(2.73)	27,41	0.201(0.056)	0.107,0.458	-0.891(0.681)	-1.802,0.095	99
		2	33.06(2.01)	28,38	0.363(0.056)	0.253,0.615	0.253(0.568)	-0.615,0.991	95
		3	32.95(2.27)	26,37	0.524(0.053)	0.389,0.646	1.398(0.794)	-0.005,2.339	86
2000/10000	None	1	50.00(0.00)	50,50	0.00(0.008)	-0.026,0.015	0.004(0.002)	0.00,0.006	100
		2	49.96(0.28)	48,50	0.00(0.008)	-0.03,0.019	-0.001(0.00)	-0.001,0.00	98
		3	49.98(0.20)	48,50	0.00(0.01)	-0.03,0.024	0.001(0.001)	0.00,0.002	99
	Middle	1	47.65(1.34)	45,50	-0.013(0.013)	-0.044,0.026	0.047(0.29)	-0.351,0.465	100
		2	48.57(1.12)	46,50	-0.027(0.016)	-0.062,0.009	-0.052(0.296)	-0.454,0.378	100
		3	48.76(1.16)	44,50	-0.042(0.02)	-0.081,0.006	-0.16(0.301)	-0.563,0.286	99
	End	1	33.00(1.90)	29,37	0.162(0.021)	0.107,0.206	-1.158(0.75)	-2.13,0.011	98
		2	33.58(1.80)	30,38	0.329(0.028)	0.26,0.396	0.015(0.565)	-0.835,0.756	97
		3	33.01(1.70)	28,38	0.493(0.033)	0.407,0.566	1.177(0.737)	-0.058,2.069	91

In comparison to the other two conditions, effects were more pronounced under the field test at the end of the test condition and these effects compounded across years. For example, the average equating constants in the 500/2500 condition were .201, .363, and .524 for Years 1 to 3 respectively. Similarly, the average difference between true and estimated TCCs under this condition across the three years were -0.891, 0.253, and 1.398, respectively, for the 500/2500 sample sizes.

Figures 5 through 10 indicate conditional differences between true and estimated TCCs (i.e., means and standard deviations of differences across iterations at each true ability level) for each of the three years of the simulation, by condition and sample size combination. Figures 5 and 6 summarize the no position effects condition and, as expected, reveal no noteworthy differences between the true and estimated TCCs.

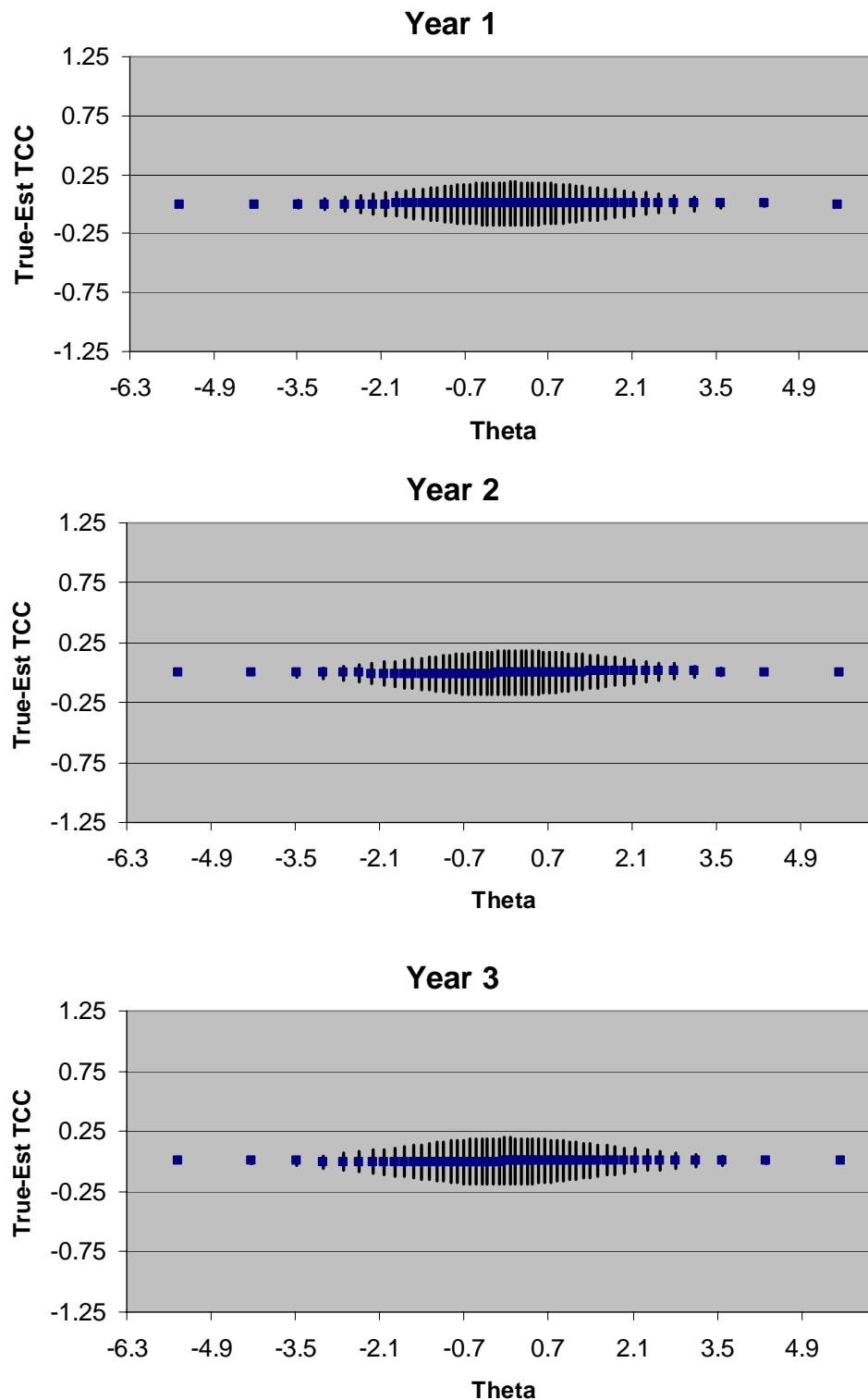


Figure 5. True Minus Estimated TCCs by Ability Level with No Item Position Change Effects Simulated by Year of Equating for the 500/2500 Condition

Note: Plotted points represent the mean difference between the true and estimated TCC and lines extending from the points represent one standard deviation, each direction, of these differences

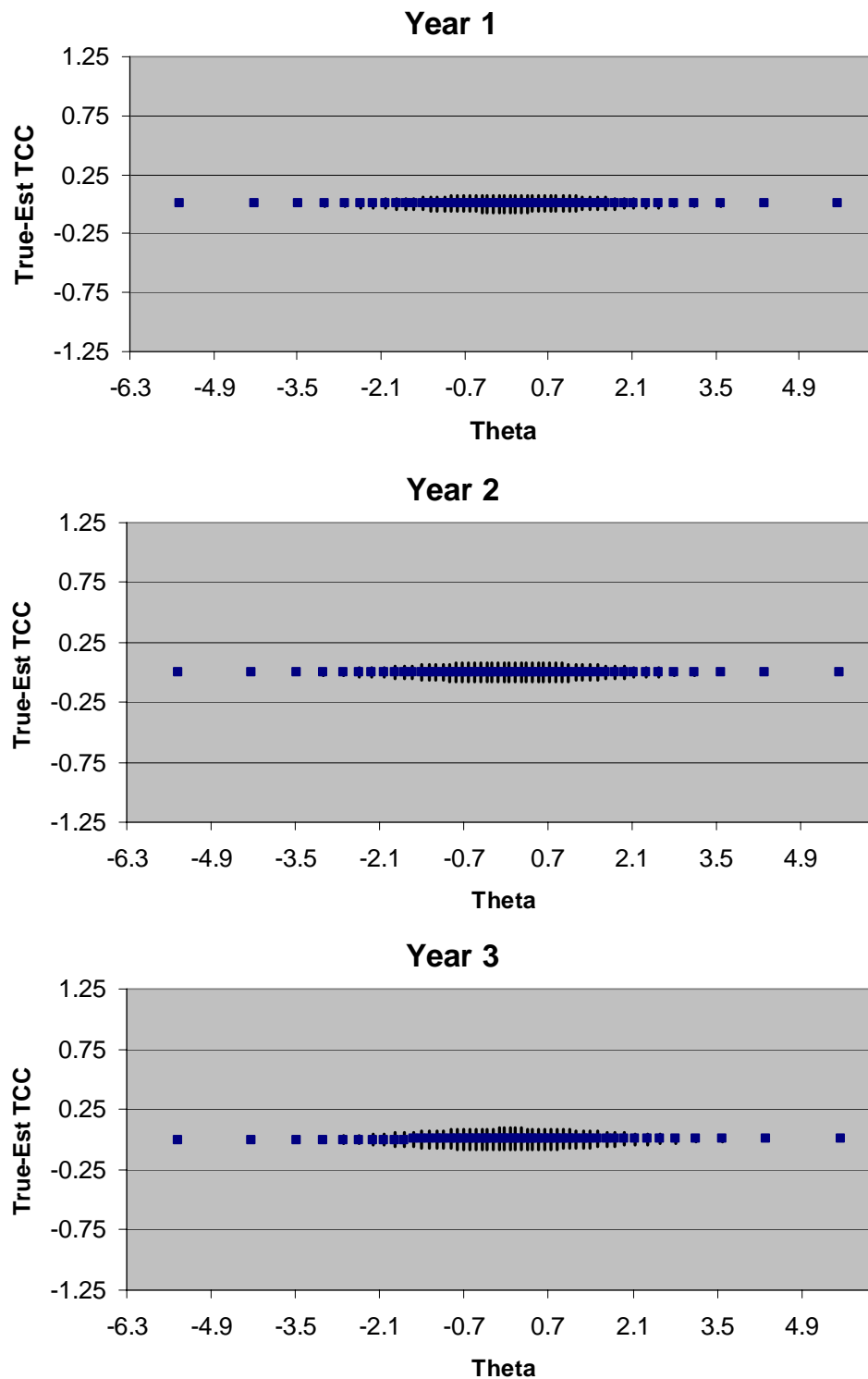


Figure 6. True Minus Estimated TCCs by Ability Level with No Item Position Change Effects Simulated by Year of Equating for the 2000/10000 Condition

Note: Plotted points represent the mean difference between the true and estimated TCC and lines extending from the points represent one standard deviation, each direction, of these differences

Figures 7 and 8 summarize the field test in the middle condition, and suggest somewhat complex results. Also provided in these figures are the conditional true minus adjusted TCC values, where the adjustment was made to the true item parameters for position effects alone (no perturbing or post-equating). This additional plot is represented by the white dots and can be thought of as the expected true student performance on the operational test. Note that the differences in Figures 7 and 8 are not centered about zero, and the direction of the differences changes after Year 1. In Year 1, the largest positive difference was .483 for an ability level of 1.663, while the largest negative difference was -.326 for an ability level of -1.688. For the prototypical form used as a basis for the simulations, an ability level of zero was located between raw scores 24 and 25. The differences between true and estimated TCCs changed from negative to positive at this point as well. Therefore, the post-equated Year 1 form over-estimated the ability of lower achieving examinees and under-estimated the raw score of higher achieving examinees. Each year of equating added to the magnitude of these effects, but the overall test appeared harder each year. The white dots in Figures 7 and 8 indicate that the item position effects resulted in true performance patterns that were similar to those reflected in the test equatings. However, the white dots appear above the mean difference between true and estimated TCCs, which suggests a small equating bias that in this case would disadvantage test-takers.

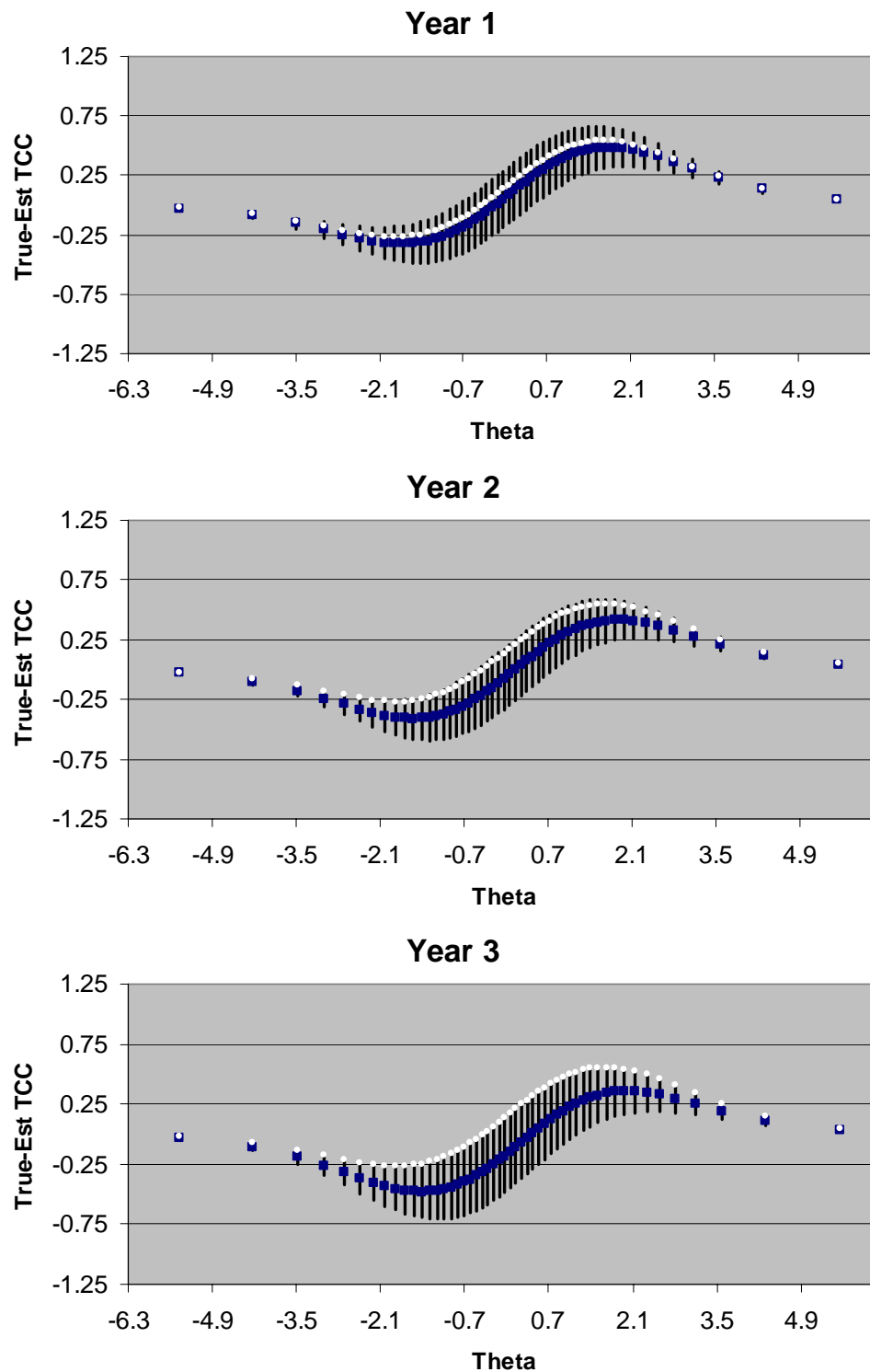


Figure 7. True Minus Estimated TCCs by Ability Level with Item Position Change Effects Simulated and Middle of Test Field Testing by Year of Equating for the 500/2500 Condition

Note: Plotted black points represent the mean difference between the true and estimated TCC and lines extending from the points represent one standard deviation, each direction, of these differences. White points represent the difference between true and adjusted TCC.

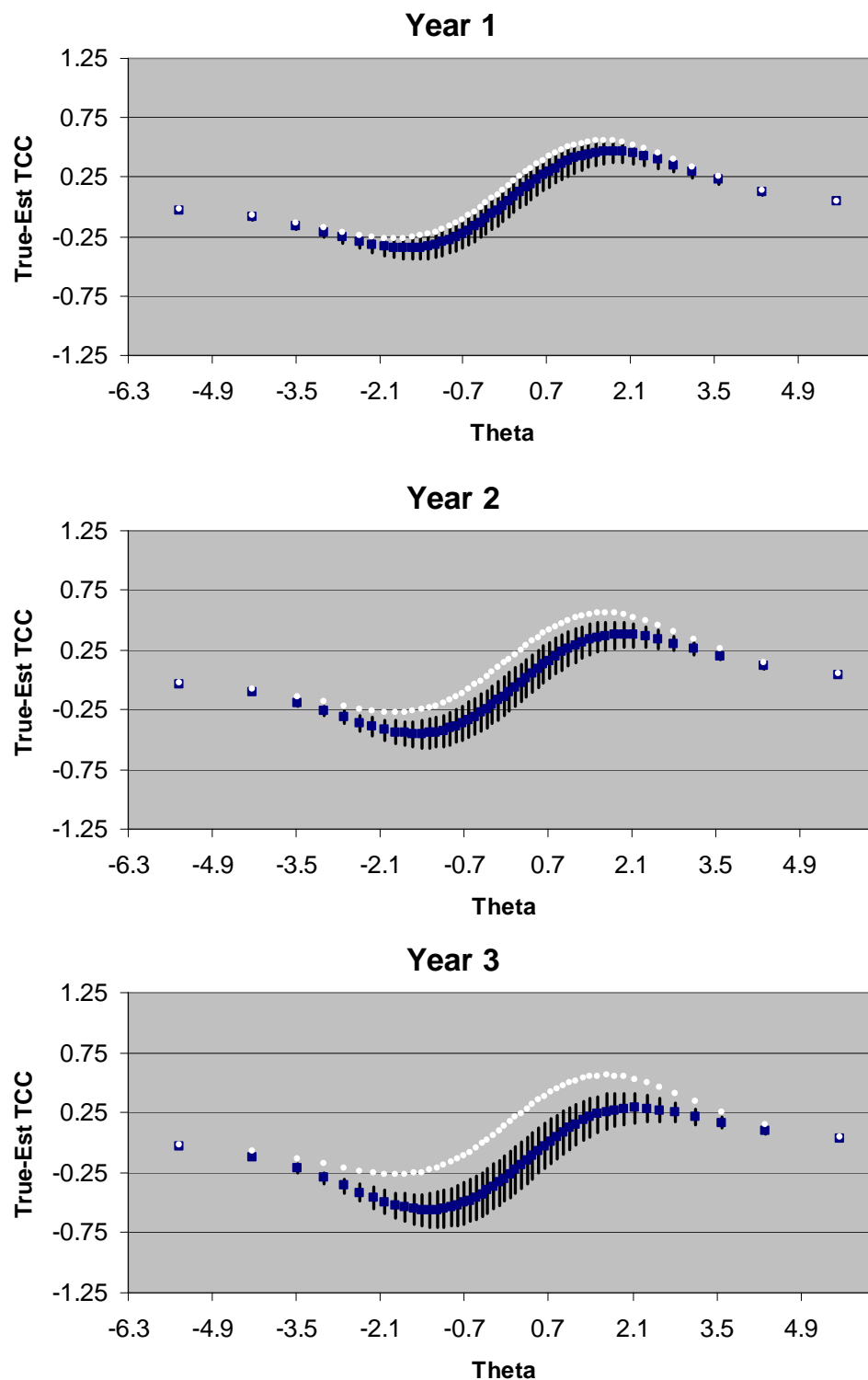


Figure 8. True Minus Estimated TCCs by Ability Level with Item Position Change Effects Simulated and Middle of Test Field Testing by Year of Equating for the 2000/10000 Condition

Note: Plotted black points represent the mean difference between the true and estimated TCC and lines extending from the points represent one standard deviation, each direction, of these differences. White points represent the difference between true and adjusted TCC.

Figures 9 and 10 summarize the field test at the end condition, and fairly dramatic effects can be seen in these plots. Note that when items are ordered generally from easy to hard the easier items are expected to be more impacted by position change effects than harder items. An easy item may move from position 60 to 1, 59 places. But a hard item may move only from 51 to 50, or 1 place. Note that these differences were also not centered about zero. In Year 1, the largest positive difference between true and estimated TCCs was .0095 for an ability level of 3.140, while the largest negative difference was -1.802 for an ability level of -1.154. The differences between true and estimated TCC changed from negative to positive between raw scores of 42 and 43, which was well above the place where ability values changed from negative to positive. Therefore, the post-equated Year 1 form over-estimated the ability of most examinees. Each year of equating added to the over-estimation, with the raw score representing the ability 1.40 dropping systematically after Year 1: 38, 37, to 36 for Years 1 to 3 respectively.

The white dots representing the conditional true minus adjusted TCC values indicate that, as expected, the tests are truly easier when the position effects are accounted for. However, the post-equating process adjusts these effects away. Conceptually, the effects of item position change are contained in the equating constant, and applied to new items upon post-equating – step 7 in our simulation process. As years increased the equating constants increased in magnitude, which resulted in a shift in the item pool, and eventually, difficulty in building a new operational form. This was dramatic observed when items were field tested at the end of the test. Recall that the operational form was randomly selected from the field test pool according to item difficulty specifications. The pool was divided into 14 groups based on .30 theta units. For example, group 1 was all items less than or equal to -2.1. The “specifications” required some representation from each; group 1 was 1 item. When field testing at the end of the test, the equating constant exceeded .30 beginning at Year 2. Adding this constant to the new field test items resulted in shifting most items in group 1 to group 2, and all items in groups 2 through 13 to the next higher group. Only a few extremely easy items remained in group 1. As the effects compounded, whole difficulty groupings became eliminated, and test forms could not be built to specifications. Of the 100 replications in Year 3, only 86 were left with one item in group 1, therefore, 14 replications were not able to build a test in Year 3. Although a small number of tests were dropped in earlier years due to the perturbing of items, most Year 3 losses were due to the effect of post-equating from Year 2 on the final item difficulties. We attempted to continue the simulation into Years 4 and 5, however, the affect of post-equating reduced the ability of the item pool to support test building, and not enough iterations resulted in successful builds to report equating constant or TCC comparison averages

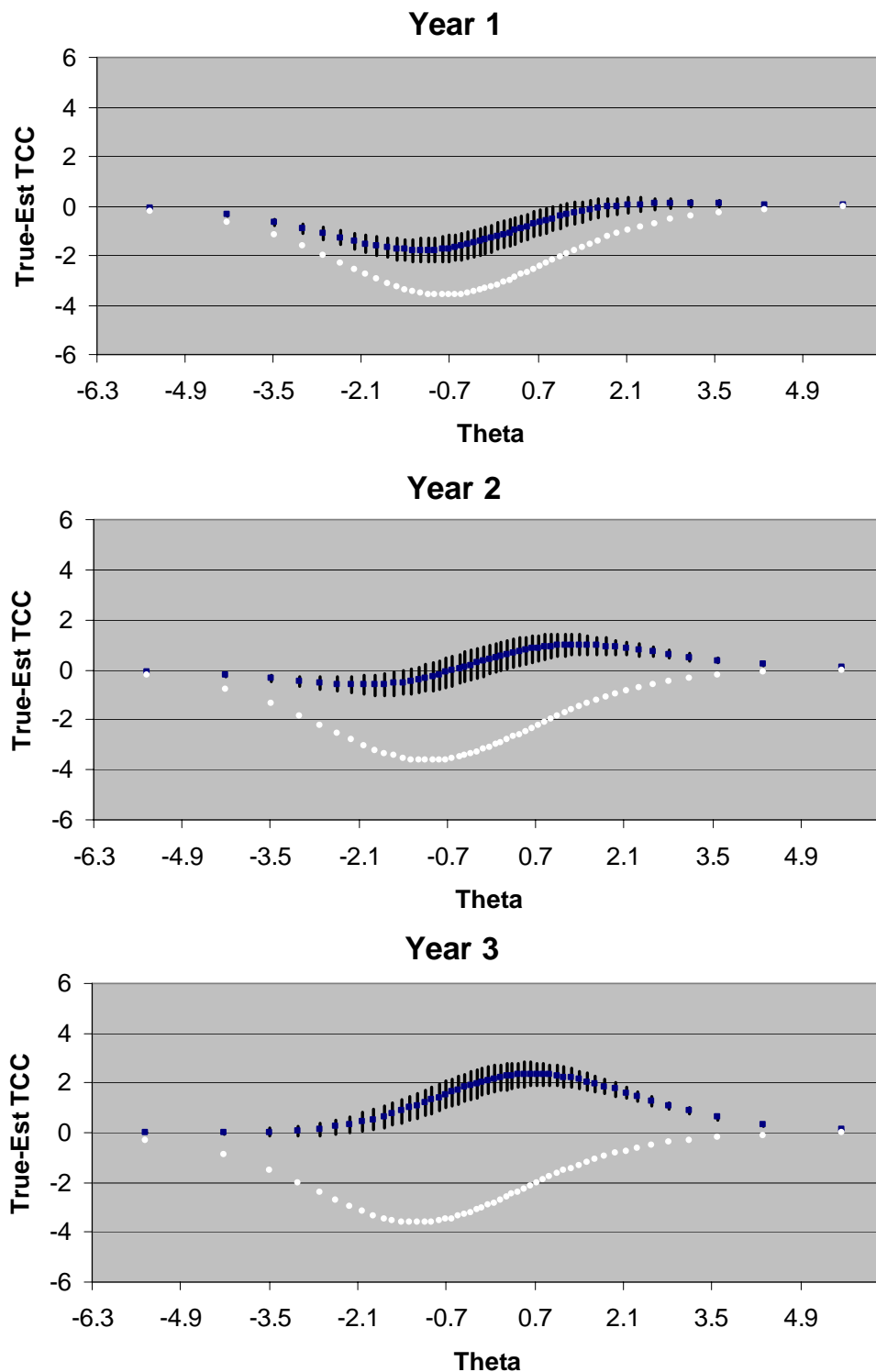


Figure 9. True Minus Estimated TCCs by Ability Level with Item Position Change Effects Simulated and End of Test Field Testing by Year of Equating for the 500/2500 Condition

Note: Plotted black points represent the mean difference between the true and estimated TCC and lines extending from the points represent one standard deviation, each direction, of these differences. White points represent the difference between true and adjusted TCC.

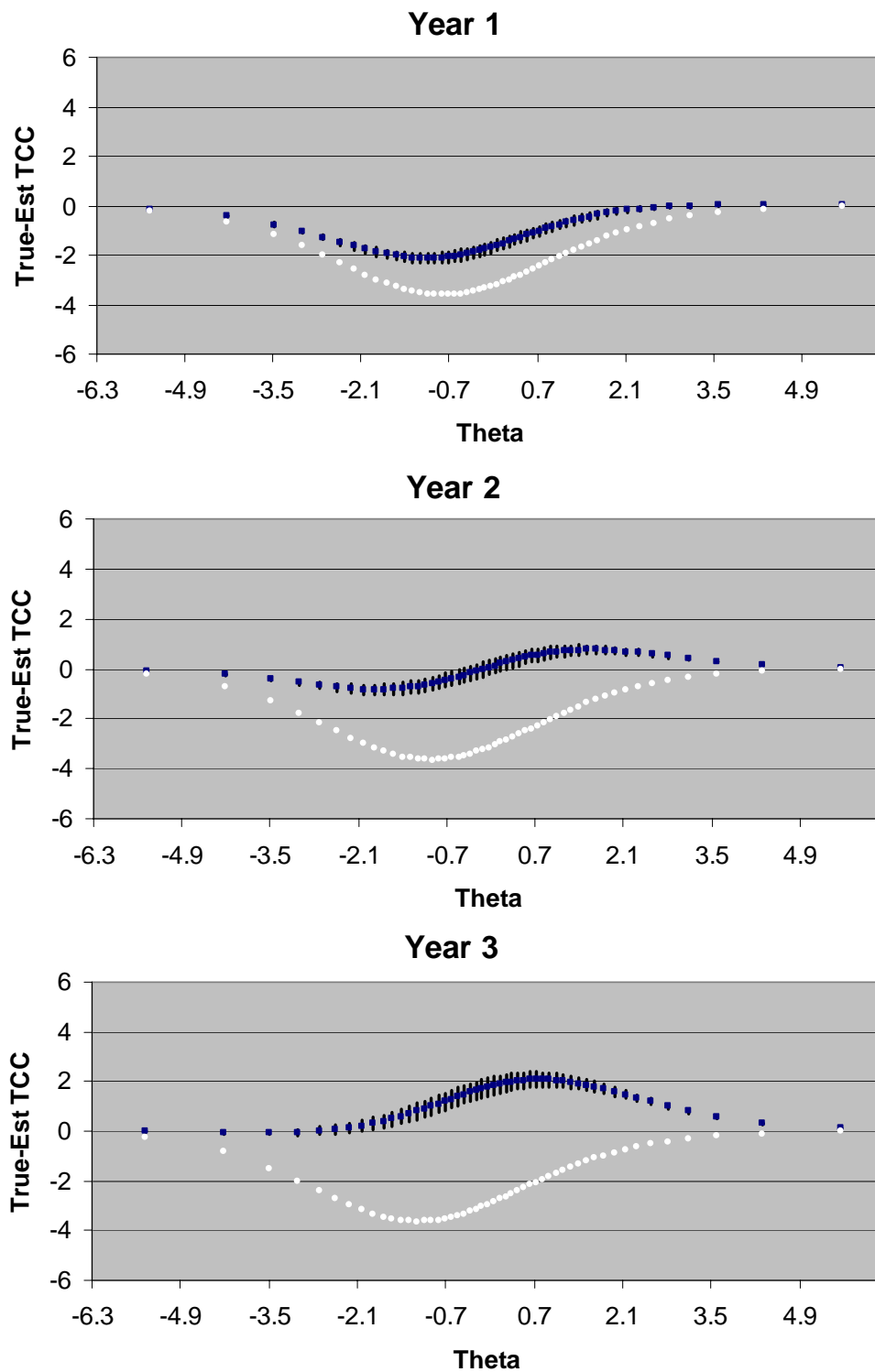


Figure 10. True Minus Estimated TCCs by Ability Level with Item Position Change Effects Simulated and End of Test Field Testing by Year of Equating for the 2000/10000 Condition

Note: Plotted black points represent the mean difference between the true and estimated TCC and lines extending from the points represent one standard deviation, each direction, of these differences. White points represent the difference between true and adjusted TCC.

We were initially puzzled by the pattern in the year 1 plots in Figures 9 and 10 because the true minus estimated TCCs suggest that the equated tests were easier than the true tests. We had expected that the equating process would result in these differences to be centered near zero in year 1. However, in the field test at the end condition, fairly large numbers of items were dropped from the equatings through the stability check (see Table 6). These tended to be the easiest items in the test. These items became much easier because of large changes in position, but they did not contribute to equating. So the estimated TCCs reflected an easier test compared to the true TCCs in year 1. In year 2, positive equating adjustments were applied to the field test items from year 1, so that the difficulty estimates for these items reflected the bias from the item position effects in the operational test. By year 3, these effects compounded to the point where the maximum differences between the estimated and true TCCs were more than two points. Thus, the tests that were truly easier appeared progressively more difficult across years.

Discussion

Choice of field test design should be based on the intended uses of item statistics information. We have offered several questions that test developers should consider when choosing among various designs. The influx of custom testing brought on by No Child Left Behind has caused test developers to recognize and operate in more complex testing environments. Included in the considerations for test developers are strategies for creating item pools with item statistics that reflect expectations when the items are used operationally, and increasingly, for creating pre-equated test forms. The discussion we provided in this paper was intended to give policy-maker and testing professionals a better understanding of the issues that are involved in such considerations. These impinge on initial field testing and test construction efforts as well as ongoing efforts involving either imbedded or standalone field tests.

One option that a test developer might choose is the creation of a *calibrated item pool*. Aspects of creating and maintaining these pools are not well described in literature. The simulation strategy presented here was to demonstrate the relative effectiveness of field testing at two locations in the test for maintaining *calibrated item pools*. Included in the simulations were random estimation error, item position effects and field test location. When item position change effects are present and operational tests are ordered from easy to difficult, field testing exclusively at either the middle or the end of the test impacts equating over time. When field testing at the middle of the test, equated test forms over-estimated the ability of lower achieving students, and under-estimated the ability of higher achieving students. When field testing at the end of the test, equated test forms first over-estimated the ability of most students, but each succeeding year, began to under-estimate the ability of students. For both field test positions, the relationship between the true and estimated TCCs were systematically different than the

relationship between the true and adjusted TCCs, where adjustment to the true values was made exclusively for item position effects (no estimation error or post-equating). In the case of field testing in the middle, the estimated TCC values were systematically higher than the values adjusted exclusively for item position effects, although these differences were small. When field testing at the end, estimated TCC values were systematically lower than the adjusted values, and the effect was dramatically compounded over time. When field testing at the end of the test, the difference between the adjusted and estimated TCCs exceeded four raw score points in some places. In testing programs with high stakes for students, these effects may impact a meaningful percentage of students. In NCLB testing programs, the effect of field testing at the end of the test would systematically -- and artifactually -- increase the percentage of students classified as proficient over time.

A meaningful finding from a pool development and maintenance perspective is that, over time, item position change effects of the magnitude we found would result in item pools that would not support building new tests to item difficulty targets. Since, over time, items of the same content and true difficulty would be equated to be more difficult, there would be increased demand to develop items that were truly easier and easier to obtain the same item difficulty distribution. This would quickly become untenable.

Our simulation example is limited by many factors, including use of the regression function to estimate the magnitude of the item position change effects and placement of field testing items in blocks. In practice, test developers may be able to use dispersed field testing designs such that position changes can be constrained. We also adopted the prediction equation for item position effects recently found by Meyers, Miller, and Way (in press); this equation may or may not generalize to other tests and operational settings. Additional research should be carried out on the impact of position effects on equating in traditional and new settings. This research should be extended to tests containing mixed item formats as the effects for constructed-response items may be different from traditional multiple-choice items.

References

- Achieve, Inc. (2007). Closing the expectations gap 2007: An annual 50-state progress report on the alignment of high school policies with the demands of college and work. Available at: <http://www.achieve.org/files/50-state-07-Final.pdf>.
- Holland, P.W., & Rubin, D.B. (1982). Test equating. New York: Academic Press, Inc.
- Holland, P.W., & Dorans, N.J. (2006). Linking and equating. In R.L. Brennan (Ed.), Educational Measurement (Fourth edition). Westport, CT: American Council on Education and Praeger Publishers.
- Haladyna, T.M. (2004). Developing and validating multiple-choice questions (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.
- Kolen, M.J., & Brennan, R.L. (2004) Test equating, scaling, and linking: Methods and practices (2nd Edition). New York: Springer.
- Meyers, J.L., Miller, G.E., & Way, W.D. (in press). Item position and item difficulty change in an IRT-based common equating design. Applied Measurement in Education.
- Meyers, J.L., Miller, G.E., and Way, W.D. (2006). Item position and item difficulty change in an IRT-based common item equating design. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.
- Schmeiser, C.B., & Welch, C.J. (2006). Test development. In R.L. Brennan (Ed.), Educational Measurement (Fourth edition). Westport, CT: American Council on Education and Praeger Publishers.