Effects of Misbehaving Common Items on Aggregate Scores and an Application of the Mantel-Haenszel Statistic in Test Equating

CSE Report 688

Michalis P. Michaelides
Independent Researcher

April 2006

Center for the Study of Evaluation (CSE)
National Center for Research on Evaluation,
Standards, and Student Testing
Graduate School of Education & Information Studies
University of California, Los Angeles
Los Angeles, CA 90095-1522
(310) 206-1532

Project 3.5 Methodological Issues in Accountability Systems/Strand 1
Edward Haertel, Project Director, CRESST/Stanford University School of Education

The work reported herein was supported under the Educational Research and Development Centers Program, PR/Award Number R305B960002, as administered by the Office of Educational Research and Improvement, U.S. Department of Education.

The findings and opinions expressed in this report do not reflect the positions or policies of the National Institute on Student Achievement, Curriculum, and Assessment, the Office of Educational Research and Improvement, or the U.S. Department of Education

EFFECTS OF MISBEHAVING COMMON ITEMS ON AGGREGATE SCORES AND AN APPLICATION OF THE MANTEL-HAENSZEL STATISTIC IN TEST EQUATING[1]

Abstract

Consistent behavior is a desirable characteristic that common items are expected to have when administered to different groups. Findings from the literature have established that items do not always behave in consistent ways; item indices and IRT item parameter estimates of the same items differ when obtained from different administrations. Content effects, such as discrepancies in instructional emphasis, and context effects, such as changes in the presentation, format, and positioning of the item, may result in differential item difficulty for different groups. When common items are differentially difficult for two groups, using them to generate an equating transformation is questionable. The delta-plot method is a simple, graphical procedure that identifies such items by examining their classical test theory difficulty values. After inspection, such items are likely to drop to a non-common-item status.

Two studies are described in this report. Study 1 investigates the influence of common items that behave inconsistently across two administrations on equated score summaries. Study 2 applies an alternative to the delta-plot method for flagging common items for differential behavior across administrations.

The first study examines the effects of retaining versus discarding the common items flagged as outliers by the delta-plot method on equated score summary statistics. For four statewide assessments that were administered in two consecutive years under the common-item nonequivalent groups design, the equating functions that transform the Year-2 to the Year-1 scale are estimated using four different IRT equating methods (Stocking & Lord, Haebara, mean/sigma, mean/mean) under two IRT models—the three- and the one-parameter logistic models for the dichotomous items with Samejima's (1969) graded response model for polytomous items. The changes in the Year-2 equated mean scores, mean gains or declines from Year 1 to Year 2, and proportions above a cut-off point are examined when all the common items are used in the equating process versus when the delta-plot outliers are excluded from the common-item pool. Results under the four equating methods

[1] The author would like to thank Edward Haertel for his thoughtful guidance on this project and for reviewing this report. Thanks also to Measured Progress Inc. for providing data for this project, as well as John Donoghue, Neil Dorans, Kyoko Ito, Michael Jodoin, Michael Nering, David Rogosa, Richard Shavelson, Wendy Yen, Rebecca Zwick and seminar participants at CTB/McGraw Hill and ETS for suggestions. Any errors and omissions are the responsibility of the author.
Results from the two studies in this report were presented at the 2003 and 2005 Annual Meetings of the American Educational Research Association.

were more consistent when a one-parameter rather than when a three-parameter logistic model was fitted. In two of the four assessments, the treatment of outlying common items had an impact on aggregate statistics: equated mean scores, mean gains and proportions above a cut-off differed considerably. Factors such as the number of outlying items, their type (dichotomously or polytomously scored), their level of difficulty, the direction and the amount of their change from Year 1 to Year 2, and the IRT model and equating transformation fitted to the data are discussed with regards to their influence on equated summary statistics.

The differential behavior of common items can be considered as a special case of Differential Item Functioning (DIF); the two different groups that respond to a common item can be regarded as the focal and reference groups, and their performance can be compared for DIF. Study 2 applies the Mantel-Haenszel statistic (Mantel & Haenszel, 1959), which is widely used for DIF analysis, on one statewide assessment that was administered to two consecutive annual cohorts of students. Sixty-nine common items, including nine polytomous items, are analyzed first with the delta-plot method and then with the Mantel-Haenszel procedure. A scheme for flagging dichotomous items for negligible, intermediate, or large DIF takes into account both the significance of the Mantel-Haenszel statistic and the effect size of the log-odds ratio; an alternative scheme developed specifically for polytomous items utilizes Mantel's chi-square statistic (Mantel, 1963) and the Standardized Mean Difference (e.g., Dorans & Schmitt, 1991/1993). The Mantel-Haenszel procedure flagged three common items, including one polytomous, for intermediate DIF. The delta-plot identified two dichotomous items only; one of which was flagged by both procedures. Assumptions are examined and it is argued that the Mantel-Haenszel procedure is more appropriate for comparing the performance of two groups because differences in the distributions of ability of the two cohorts are taken into account. The availability of schemes that classify items according to the amount of DIF they exhibit can be informative for the judgmental decision on how to deal with flagged items. However, some caveats relating to test construction and implementation of the equating design are noted if the proposed procedures are to be applied effectively. The same common items and an adequately large number of them must be presented in corresponding forms across administrations. This is pertinent especially for assessments employing a matrix-sampling design, where the common-items are spread among many forms.

Introduction

Large-scale testing programs provide scores for individual student achievement, and aggregate scores for classrooms, schools, districts, or states. Scores are often derived from different versions of a test administered over multiple occasions. Not all examinees respond to the same test form. While this is a way to guard against the overexposure of the content and ensure the security of the test, it creates the problem of interchangeability of the scores. Alternate test forms will be differentially difficult for examinees; however for the sake of fairness, it should not matter to them which test form they take (Lord, 1980). Test equating methods are statistical adjustments that establish comparability between alternate forms built to the same content and statistical specifications by placing scores on a common scale (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, 1999; Kolen & Brennan, 2004). In the common-item nonequivalent groups design, two forms are equated through a subset of common items embedded in both forms. Performance on the common items is used to establish the linking relationship between groups taking the alternate forms.

A key assumption made when equating is performed under the common-item nonequivalent groups design is that the statistical properties of the anchor items are stable across forms; when two groups respond to two alternate forms, the common items must function similarly in both forms (Hanson & Feinstein, 1997; Wainer, 1999). If two groups of examinees respond differently to the same item, then that item might not be appropriate to be included in the equating process. If an equating item demonstrates a large change in its difficulty index or its Item Response Theory (IRT) parameter estimates, it raises suspicion, and calls for inspection. Upon inspection, analysts seek to determine possible reasons for why the item functions differentially. They can speculate whether the differential performance is related to the purpose of measurement, i.e., if it reflects a true change in the proficiency of the examinee cohorts, or if it is due to irrelevant circumstances, such as a change in the position of the item in the test form. It may be discarded from the equating item pool and treated as a regular, non-common item, as if there was no connection between the item in the first and the item in the second form. Inclusion or exclusion of an item from the equating pool is a matter of judgment and affects the equating function.

In the next section, a review of the existing literature reveals that characteristic indices of items change and item parameters are not invariant across different administrations – although IRT models that are usually fitted to test data rest on the assumption of parameter invariance. For example, variations in instruction and curricular emphasis result in items having differential difficulty for different groups of examinees. Apart from content reasons, item parameter drift might be due to context effects such as changes in item position or inadvertent changes in the precise wording or formatting of items.

IRT or classical item statistics may be used to examine whether embedded common items are functioning differentially for groups taking different test forms (Kolen & Brennan, 2004). In practice, a procedure used to examine the volatility of equating items' difficulty values is the delta-plot method, a simple and comprehensible method for studying the item-by-group interaction, which makes use of the classical test theory difficulty indices, the p-values (Angoff, 1972). It is a graphical procedure that flags outliers in a scatter plot of transformed p-values of the common items obtained from two groups of examinees. The points that lie at a distance from the "cloud" of the majority of the points represent the common items whose p-values differ by an unexpectedly small or large amount. Those items are candidates for exclusion from the common-item pool. The delta-plot method is widely implemented because it is practical and does not involve IRT calibrations, which would be the case if IRT parameters were compared, and because it provides prima-facie evidence regarding anomalous changes in item difficulties across administrations.

The first study in this report investigates if, and to what extent, decisions on whether to keep or discard the outlying items of the delta-plot from the common-item pool impact the equating transformations and the resulting score distribution summaries. The effects on the equating transformation and the equated score aggregates are examined, when all common items are used versus when misbehaving items identified by the delta-plot method are discarded from the common item pool.

Although practical, the delta-plot method is a crude procedure in the sense that it summarizes the information from an item in a single number, its p-value, and looks at how that number is related to the p-values of the remaining common items. It also transforms the p-values through an inverse normal transformation, which changes their distribution in a somewhat arbitrary way.

The problem of inconsistent behavior of common items across administrations can be viewed as an instance of differential item functioning (DIF), where the two groups taking two different forms with some items in common are the focal and the reference groups. Mantel and Haenszel (1959) proposed a common odds ratio to assess the strength of association in three-way *2x2xK* contingency tables. The ratio estimates how stable the association between two factors is in a series of *K* partial tables (the strata of a third factor.) Originally proposed by Holland (1985) as a potential method for studying DIF in any two groups of examinees, Holland and Thayer (1988) published a paper that popularized the Mantel-Haenszel statistic as a measure of DIF. It can be used to test the null hypothesis

$$H_0 : \frac{p_{Rj}}{q_{Rj}} = \frac{p_{Fj}}{q_{Fj}}$$

Namely, the odds for answering correctly item *j* for the reference group *R* are equal to the corresponding odds for the focal group *F*. Note that $q_{ij} = 1 - p_{ij}$. The alternative hypothesis is

$$H_1 : \frac{p_{Rj}}{q_{Rj}} = \alpha_j \frac{p_{Fj}}{q_{Fj}}$$

where $\alpha_j = \dfrac{p_{Rj} q_{Fj}}{p_{Fj} q_{Rj}}$ is the common odds ratio ($\alpha_j \neq 1$).

The focal and reference groups are matched on ability, the third factor, using a test score interval as a proxy. The procedure provides a chi-squared test statistic as well as an estimator of $_{j}$ across the *K 2x2* tables. The latter—one of the strengths of the Mantel-Haenszel method—is a measure of the effect size, or how much the data depart from H_0, an important feature since conventional statistical significance can be easily obtained with large enough samples.

The Mantel-Haenszel procedure may be implemented in the context of equating to identify which common items behave differentially across administrations by considering the two annual examinee cohorts as the focal and reference groups. In contrast to the delta-plot method, the proposed alternative can provide a measure of how much an item departs from consistency, by comparing examinees of similar proficiency in the two cohorts. The second study in this report describes an implementation of the Mantel-Haenszel procedure to detect differential item performance permitting more meaningful comparisons of item behavior across administrations.

In summary, this report looks at statistical problems associated with the selection of common items in the context of IRT equating of assessments under the common-item nonequivalent groups design. The following two research questions are addressed:

1. What is the effect of keeping in versus discarding from the common-item pool common items flagged as outliers by the delta-plot method on equated score summaries?
2. How can the Mantel-Haenszel procedure be applied to investigate differential item performance across test forms? How does it compare to the currently used delta-plot method?

IRT, Assumptions, and the Promise of Parameter Invariance

Test equating is a component of a larger, cyclic, measurement process that involves test development, administration, analysis, scoring, reporting, and evaluation (Hattie, Jaeger, & Bond, 1999). When a model, such as any IRT model, provides the conceptual measurement framework for this process, the results depend on how well the data fit that model.

Parameter invariance, the property of IRT item parameter estimates to remain unchanged across various groups of examinees, and ability estimates to remain invariant across groups of items, gives IRT its applicability and usefulness (Allen, Ansley, & Forsyth, 1987; Hambleton, Swaminathan, & Rogers, 1991; Linn, 1990; Lord, 1980). According to parameter invariance, if the IRT model fits the data perfectly, then parameters will be invariant across administrations, except for sampling fluctuations that introduce random error in the responses of examinees. In that case, the changes in the behavior of item parameter estimates would follow a systematic pattern depending on the changes in the size and proficiency of the different examinee groups.

IRT makes strong assumptions and its promise for invariance depends on the degree that the model assumptions, and particularly unidimensionality, hold (Miller & Linn, 1988). Violation of the unidimensionality assumption is potentially a major source of problems for IRT equating (Skaggs & Lissitz, 1986). Suppose estimation of item parameters using data from two different groups of examinees yields different item parameters, or equivalently, that two test-characteristic curves exist for the same test.[2] Examinees in the two groups who get equal ability estimates from the model would have different probabilities of answering items correctly. According to Lord (1980), if a test discriminates between examinees of the same ability, it actually measures a

[2] Because item parameters are invariant only up to a linear transformation of the ability scale, item parameter estimates obtained using different examinee groups would need to be transformed to place them on a common scale. This illustration refers to differences in item parameter estimates that remain after such rescaling has been carried out.

dimension other than the intended ability. Unidimensionality is not defensible and the assumption of invariance is then dubious. Even though unidimensional models do not fit to educational settings where multiple proficiencies are engaged simultaneously in tasks of interest, common practice employs unidimensional models to analyze educational tests.[3]

To make an argument for unidimensionality, a dominant component affecting test performance would suffice (Hambleton, et al., 1991). However, empirical findings have been consistent in pointing to departures from unidimensionality that are usually large, thus casting doubt on the underlying assumptions and the inferences drawn from the model. Differential instruction or dissimilar emphases in curricula, for example, introduce additional dimensions to the dimensions that the test is built to measure, such as the opportunity to learn specialized topics, or time lapsed since a topic was taught. If content tested by one item is emphasized by one instructional program more than another, then the item will likely be differentially difficult for the two groups (Masters, 1988). Context effects, such as the position of a common item in two test forms, are likely to introduce systematic differences in the item parameters as well (Kolen & Brennan, 2004).

Traub (1983) described thought experiments and numerical examples for conditions that would cause the assumption of unidimensionality to fail. Differences in instruction, such as differential emphasis for training in one topic versus another, would lead to differentiation of constructs, and would require more than one number, or one dimension, to describe examinee proficiencies. Similar effects would appear as a result of content-irrelevant skills, including the ability to work through items quickly under speeded test conditions, and the propensity of students to guess when they are not sure about the correct answer. Variations across students in these characteristics would introduce additional dimensions to the test. He concluded that "[n]o unidimensional item response model is likely to fit educational achievement data" (p.65); a simple model should not be expected to characterize complex constructs accurately.

Large changes or differences in instructional experience may be needed to produce practically significant violations of assumptions, and the effects may be very specific and limited to few items of a test (Linn, 1990). However, IRT models are

[3] Some researchers have argued that the use of latent trait models should be reconsidered because they suffer from serious defects and contestable assumptions (Goldstein, 1980).

approximations at best. And thus anomalous item behavior may not be ruled out, even if models perform sufficiently well to justify their continued use.

The problem becomes more profound in test equating when equating functions are derived from item parameters of the common items. It is not unusual to come across a few common items that do not follow the behavior of the majority of common items. Those misbehaving items are checked for possible reasons that caused their anomalous behavior. It is then a judgmental decision whether to keep or discard them from the equating pool. An obvious explanation might exist for that behavior, such as modified wording of a question, or differential emphasis on the content of the item, or there may be no immediate compelling reason triggering that behavior. In any case, keeping those items in the equating will place one group at a slight advantage over the other. It may be more appropriate that they be discarded from the common item pool and treated as different items in the two administrations.

Curricular Effects on Item Parameter Estimates

Many empirical studies address the adequacy of IRT models by examining whether parameter invariance or unidimensionality holds using real or simulated data. Miller and Linn (1988) examined the effect of differential instructional coverage on item characteristic functions. They clustered students who participated in the Second International Mathematics Study into curriculum clusters based on their teachers' ratings of their opportunity to learn each item during the previous year. Item characteristic curves for the arithmetic and algebra items for each of the curriculum clusters were compared. Large differences were detected between the curves, indicating that item parameters were influenced by variations in opportunity to learn.

Masters (1988) provided evidence for differential item performance caused by the opportunity to learn particular content in high- versus low-level mathematics classes. For example, items on content that one group had more opportunity to learn had different difficulty parameters when separate calibrations were made for each group. If the two groups' responses were calibrated simultaneously, the difficulty parameter would fall between the previous two values, and the discrimination parameter would be higher if the group that had more opportunity to learn was on average of higher ability.

Content analysis may help explain findings of item parameter drift (Linn, 1990). Bock, Muraki and Pfeiffenberger (1988) found differential linear drift of the item location parameters in items of a College-Board Physics Achievement Test over ten years. They associated the direction of the drift with the content of the items in a pattern that reflected a changing emphasis in secondary school physics curricula. Item parameter drift is more likely in subjects that change easily over time. Considering a lack of substantial drift in English items over that same time period, they attributed the noticeable drift in physics items to the greater likelihood of change in physics curricula. Among 29 mechanics items, 11 that referred to basic concepts became easier over time, while the difficulty of 10 other items less related to basic concepts increased. Their evidence suggests a decreased emphasis on advanced and specific topics, which may reflect a back-to-basics approach in physics textbooks. A pair of mechanics items on the difference between mass and weight, one of which used metric and the other English units, exhibited drift in opposite directions. The cases moved in a direction that reflected the introduction of metric units at the end of the 1970s. Apart from systematic item location drift, Bock, et al. (1988) observed occasional anomalies in some items. They suggested that such cohort-specific effects are unexpected in large nationwide samples, but may reflect special attention given to some topics by the media or publications accessible to physics teachers.

Sykes and Fitzpatrick (1992) classified a large number of items from consecutive administrations of a professional licensure examination into four content categories. In one of the four categories, they detected a significantly greater drift of Rasch *b* parameter estimates. They hypothesized that the "differential change in *b* values is attributable to shifts in curriculum emphasis, with the most pronounced shift occurring for the content covered in this category" (p.210).

A series of studies by Mehrens and Phillips suggested that differential instruction and textbook effects are not a serious concern for test validity. Neither the different textbook series used in Grades 3 and 6 for reading and mathematics, nor the degree of instruction-test match based on teachers' ratings were found to impact standardized test scores significantly (Mehrens & Phillips, 1986). The small impact of these two curricular factors on unidimensionality was evaluated by a factor analytic method: the percentage of variance for the large first factor did not change noticeably, and the second factor remained relatively small across groups with different curricula (Phillips & Mehrens, 1987). In a third paper, they reported that item p-values and Rasch difficulty parameter estimates were similar across student groups using different textbook series (Mehrens & Phillips, 1987). The authors list a number of potential reasons for the lack of curricular impact, including the lack of power to detect differences, the precision of teachers' ratings of instruction-test match (Phillips &

Mehrens, 1987), and emphasis on general competence versus specific details related to curricular objectives. Linn (1990) further comments on the findings by Mehrens and Phillips that their studies were done in elementary grades with widely used textbooks, in contrast to studies in higher grades that have qualitatively more different instructional experiences, and which found a demonstrated impact on test performance.

Much research on item parameters emerged from studies of customized tests and the validity of estimates drawn from actual or simulated customizations. In the early 1980s, national tests were often customized by local authorities and adjusted to extract national normative scores for the local examinees. The validity of such inference has been questioned. Consistent findings indicate that item calibrations are not invariant across samples. Yen, Green, and Burket (1987) present a case where the IRT difficulty parameter estimates in the national calibration of a mathematics test changed systematically at the local level; in a local calibration the measurement items were relatively more difficult, while the numeration items were relatively easier, suggesting that different local curricular and/or instructional characteristics influenced parameter estimates.

Allen, et al. (1987), Linn (1990), Way, Forsyth, and Ansley (1989), and Yen, et al. (1987) provide examples of the effects that customized tests non-representative of the original tests can have on ability estimates. Tests customized by selecting specific content areas gave systematically higher ability estimates than estimates based on the full test; the same was not true when content was sampled representatively. The magnitude of the overestimates seemed dependent on the number of items deleted from the full test (Way, et al., 1989).

Context Effects on Item Parameter Estimates

Apart from the content of items, the context in which they are presented also influences the estimates of item parameters. “A context effect occurs when a change in the test or item setting affects student performance” (National Research Council, 1999). Masters (1988) considers (a) opportunity to answer, due to speeded tests, and fatigue, and (b) test wiseness as sources of differential item performance reflected in item parameters. The two factors pertain to different examinee groups and introduce measurement disturbance. Items that appear at the end of a test and items that are sensitive to test wiseness skills will favor students of higher ability and thus produce inflated discrimination parameters. Had the items appearing at the end of the test been presented earlier, they would have been attempted by more examinees and would have likely exhibited lower discriminations.

Kingston and Dorans (1984) examined location effects for 10 item types on the GRE test. They found changes in the IRT equatings when those items were moved to a different position in the test. The effects were larger for the Quantitative subtest than for the Verbal, and even more profound for the Analytical. Practice and fatigue effects clearly depended on the location of an item and seemed to interact with the type of the item (analytical, quantitative, or verbal).

Yen (1980) reports that the location of an item in a booklet frequently affected the value of its difficulty parameter. Items placed at the end of a test had higher parameter estimates than when presented at the beginning. Item location only partially explained parameter change in her paper. Similarity of item arrangements seemed to be another factor. The booklets "with the most similar item sequences tended to have more strongly related item parameter estimates than the booklets with the least similar item sequences" (p.308). Discrimination parameters seemed to be influenced slightly by the number of items being calibrated. Yen did not, however, find systematic patterns for discrimination parameters. In contrast, Sykes and Fitzpatrick (1992) reported that changes of item location parameters were unrelated to changes in the booklet or the test position, and item type (tryout or scorable).

A requirement for sound equating is that the equating function must be population invariant; the choice of (sub) populations to estimate the equating relationship between two tests should not produce large discrepancies (Dorans & Holland, 2000). Performance on items (common items when the common-item nonequivalent groups design is used) drives equating functions. Differential item performance between groups would cause dependency of the equating relationship on the population. In a study of traditional equating methods, Kingston, Leary, and Wightman (1985) looked at the equating functions between subgroups that took two different forms of the GMAT. Equated scores derived from the male and female subgroups were very similar, as were those derived from age or random subgroups. In a comparable study on the GRE and sex, race, field of study, level of performance and random subgroups, Angoff and Cowell (1985) also found support for the population invariance of equating (both studies reported in Cook and Petersen, 1987).

A third study by Cook, Eignor and Taft (1988) reached different conclusions. What was special about this study was that the samples employed to generate the equating transformation did not come from the same test administration, but from fall versus spring administrations of the test, and thus subgroups used to link the tests were dissimilar. Cook, et al. (1988) demonstrated that when curriculum-related biology achievement tests were given to groups of students at different points in time after learning the content, item parameter estimates were unstable. For instance the correlation coefficient between the delta values[4] of 58 common items in two consecutive fall administrations was 0.99 as opposed to 0.79 between the fall and the spring administration values. Groups taking the test at different points in their coursework could not be considered as samples from the same population; recency of instruction, the time lapsed from when the material was taught, appeared to influence item parameter estimates. In addition, both linear and nonlinear, including IRT, equating methods were not robust in such cases, giving very disparate scaled score summary statistics. In contrast, the statistics from equating forms administered at the same time period in consecutive years, i.e., fall of the first year and fall of the second year, were similar under all equating methods.

Disclosure of, or familiarity with items is another potential cause for changes in item location parameters. A security breach could have unpredictable effects on equated scores depending on whether the items exposed are common or not, and on the magnitude of the breach (Brennan & Kolen, 1987). A study simulating increasing levels of anchor item disclosure by randomly selecting and changing incorrect to correct responses resulted in an increasing drift in difficulty parameters as disclosure moved from low to moderate levels (Mitzel, Weber, & Sykes, 1999). More importantly, even at modest exposure levels, IRT equated score distributions altered considerably. Under a traditional Tucker linear equating method, Gilmer (1989) found modest effects on the passing rates on a certification test due to simulated item disclosure.

Desirable Characteristics of the Common Items

Common items provide the statistical means for equating test forms and making scores from different administrations of the same testing program comparable. Since tests need to be built to the same content and statistical specifications for the comparisons to be meaningful, the anchor items should proportionally reflect the specifications of the total test if they are to reflect group differences adequately (Brennan & Kolen, 1987; Cook & Petersen, 1987; Kolen & Brennan, 2004). For a

[4] Delta values are transformed proportion correct values defined as the inverse normal transformations of the p-values rescaled by multiplying by -4 and adding 13.

nonrandom, common-item equating design Budescu (1985) noted that a high correlation between the anchor subtest and the two total tests is a necessary condition for stable and precise equating. Klein and Jarjoura (1985) argued that "it is important that the common items directly reflect the content representation of the full test forms. A failure to equate on the basis of content representative anchors may lead to substantial equating error" (p.205). In contrast, in a simulation study, IRT equating was fairly robust to violations of the representativeness of the set of common items (Béguin, 2002); although the equating was based on a unidimensional model, with the addition of a second dimension in the non-common items, it performed at least as well as a multidimensional model.

Adequate numbers of common items need to be included to reduce random equating error, particularly in educational tests, which are not strictly unidimensional. As a rule of thumb, at least 20 items or 20% of the length of a moderately long test should be used as anchors (Angoff, 1971; Kolen & Brennan, 2004). The longer the anchor test, the more reliable the equating will be (Budescu, 1985). Additional precautions to avoid systematic influences on anchor items relate to their positioning, which should be approximately the same in the alternate forms (Cook & Petersen, 1987), and their presentation, which should be identical, i.e., without changing the text (Cassels & Johnstone, 1984) or the order of multiple-choice options (Cizek, 1994). The researchers making these caveats have shown, as have most of the studies previously reviewed, that performance on items is sensitive to such variations.

Before they are judged appropriate for the equating process, anchor items must pass additional analyses after the administration of the tests to scrutinize their behavior, as reflected in item parameter estimates. Items that behave consistently over multiple administrations are appropriate for use in the test equating process. Items indicating anomalous parameter changes over time are rejected from the linking item pool and treated as regular, non-common items. Hence, the number of anchors in the test should be sufficiently large to effectively complete the equating task after the rejection of some items.

IRT is utilized to study the performance, and hence the appropriateness, of the linking items. A common approach is to scale two to-be-equated forms separately. Each calibration yields item parameter estimates that are used to generate a transformation to place the tests on a common scale. For example, a scatter plot of the common items' difficulty parameters estimated in the calibration of the first form versus that of the second would show an approximately straight line under a satisfactory IRT model fit.

Some random variation is expected, but clear outliers would suggest that the assumptions of the model are not met. On a second scatter plot of classical item difficulties of the common items, any outliers, i.e., items with difficulty indices differing notably in the two forms, are likely to be excluded from the common item pool and considered as non-common items in the respective test forms.

Studies on the effect of simulated item parameter drift, i.e., the differential change in item parameters over time, on estimates of examinee proficiency have shown that individual ability estimates are fairly robust to non-common item parameter drift. When drift contaminates the common-item pool ability, estimates are influenced. Stahl, Bergstrom, and Shneyderman (2002) simulated increasing levels of item parameter drift and observed the impact on Rasch estimates of examinee measures. Under conditions of simulated increase in difficulty of various groups of items, and by varying the number of drifted items and the direction of the drift, ability estimates were robust. By setting a pass/fail cut-score, the majority of the misclassifications were within the 95% confidence band of the cut-score, "indicating that the misclassifications may be due purely to error of measurement and not to the effect of drift" (p.8). Wells, Subkoviak, and Serlin (2002) applied a 2PL IRT model, and in addition they examined what the effect was on ability estimates when drifted items were excluded from equating; they found little effect. Fitting a 3PL model, Huang and Shyu (2003) simulated conditions of drift in the discrimination and difficulty parameters, varied the sample sizes and the percentage of the common items with drifted parameters and performed equating with or without the drifted items. When drifted items were excluded from equating, the equated scores did not differ much from the baseline scores; they did affect mean and passing scores when they were kept in the common item pool. Large increases in the difficulty parameters and when the drifted items constituted half of the common item pool had the more profound consequences, especially with a small sample size of 500.

The above studies of impact of item drift on estimates of ability—whether equated or not—were all simulation studies, modeling usually unrealistically large drift in item parameters, typically in one direction and for a large number of items (common or not.) In real data, there are only few common items that demonstrate large changes in their item parameters. And those changes are not unidirectional; an item may become easier while two others become more difficult, thus partially negating some of the effect of item drift. Therefore, in real situations the effects of change in estimated parameters will probably be much less. It remains true though that the effect, if any, on aggregate scores will still be much more profound than on individual examinee scores, and may be considerable.

Implications of Dealing With Common Items With Anomalous Behavior

The analysis of linking items to determine their appropriateness for use in equating gives rise to a number of concerns regarding the valid interpretation and use of equated test scores. Stable performance of students on tests is not necessarily a desirable property. Educational systems expect and seek progress in their students' learning, not only as a result of normal educational practice, but also through the implementation of innovative programs, reallocation of resources, new policies and reforms in the curriculum or administrative procedures. Annual cohorts of students with different educational experiences cannot be regarded as randomly equivalent – hence, a nonequivalent groups design with common items is appropriate for equating their scores. If an accountability system successfully encourages the reallocation of instructional resources, then some common items answered by both groups could appear anomalous possibly because they are indicating real effects: that the reform initiative has indeed made a difference and performance on the relevant items has changed. Consequently, their classical and IRT parameter estimates change as well. When the items reflecting the results of the reallocation are removed on statistical grounds because of presumed violations of model assumptions, such as parameter invariance, the effects of the reform may be adjusted away.

There are reasons to believe that the measurement practice outlined above pertains to the validity of the test; the soundness of interpretations given to the equated test scores is jeopardized. "If items that are found to be most sensitive to instruction are eliminated so that the IRT assumptions are better satisfied, there is a real danger that IRT will do more to decrease than to increase the validity of achievement test scores" (Linn, 1990, p.136). If items examining certain curricular domains – and particularly those domains at which recently implemented policies would be targeted – are deleted from the linking pool, the content domain that the test is constructed to measure is redefined in ways that cannot be determined and limited to those items that do not disturb the model's assumptions. The removal of anomalously behaving common items, which capture a change in emphasis in a domain, is in discordance with the value placed by the system on that particular domain. As with the process of construction of good tests, items cannot be chosen merely on the basis of their psychometric attributes.

Content would be a legitimate consideration in deciding which items remain in the anchor pool.

Test score gains or drops are pervasively used for evaluations of innovative programs and policies as well. But the validity of test scores as indicators of purposive change becomes contestable since, ironically, the equating practice might remove the effects of curricular changes the testing program purports to evaluate.

Study 1: Effects Of Delta-Plot Outliers On Equated Score Summaries

Overview

Consistent behavior as denoted by item indices is a desirable characteristic for the common items in test equating. The delta-plot method is a practical procedure that identifies common items with anomalous changes in their item difficulties across two administrations, compared to the rest of the common items; these misbehaving items are flagged as outliers on a scatterplot and are likely to drop to a non-common item status.

The outliers need to be inspected to determine plausible causes for the observed inconsistencies. Even though it may be difficult to identify content or context effects that have influenced item performance with certainty, considering a range of plausible explanations can help to decide whether a misbehaving item should be discarded from the common-item pool or not. If the content of an outlying item strongly suggests that it has captured a "real" change in examinee proficiency on the measured construct, then that item might still be kept in the common-item pool. How to deal with the delta-plot outliers is not simply a statistical issue. The final decision involves judgment. Study 1 looks at what happens to equated score distributions when the outliers are either kept in or discarded from the common-item pool. It examines how the equating would vary depending on whether outliers were included in the common-item pool or not, without trying to determine why they behave as outliers.

The data sets and the methodology used in Study 1 are described in the following sections. The results section presents the effects of outliers on equated mean scores, mean gains or declines, and proportions above a cut-off point, followed by a discussion of the findings.

Data Sources

For the first study, data from four statewide assessment programs from three states were analyzed. For confidentiality reasons, the names of the states cannot be disclosed, and thus will be called States 1, 2, and 3. These assessments were administered to different grades and the subjects included mathematics, science, and social studies. All three states had had their testing programs in place for at least two years prior to the administration of the assessments.

For each of the four assessments, there were data from two successive annual administrations, referred to as Year-1 and Year-2 assessments respectively. Each Year-2 assessment can be linked to the corresponding Year-1 assessment through a set of embedded common items.

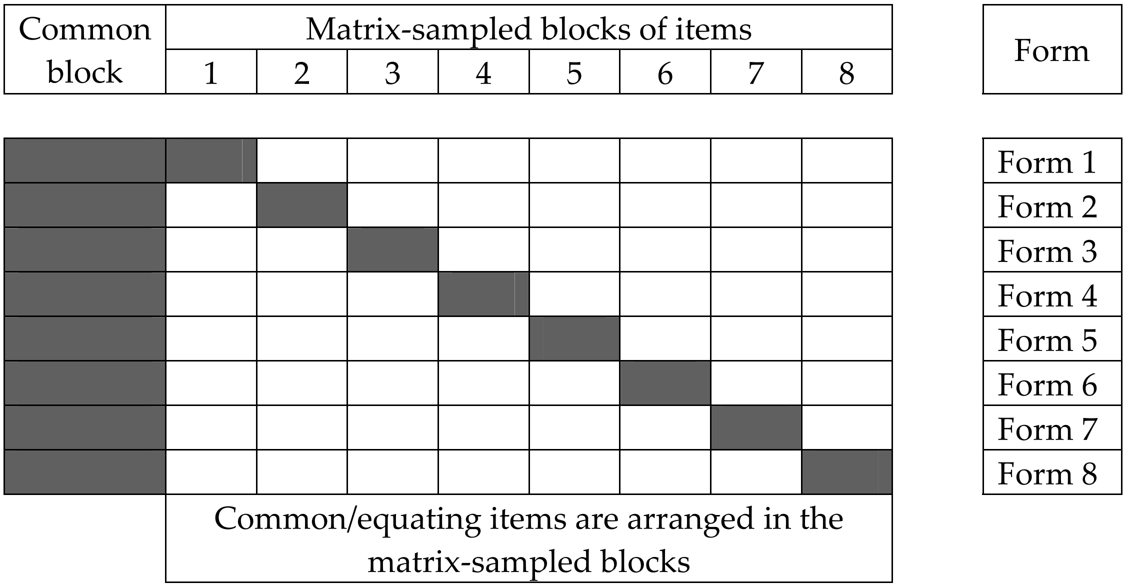
Table 1.1 gives characteristics of the assessments. The populations tested constituted the annual cohort of students graduating from the respective grade in each state and are relatively large, ranging from 7128 to 17371.

Table 1.1

Information for the Four Assessment Data Sets Analyzed

Subject	Grade	State	Number of examinees	
			Year 1	Year 2
Mathematics	8	1	7258	7128
Science	11	2	14244	14565
Social Studies	6	3	17126	17371
Science	6	3	17128	17371

A single administration at the end of a school year consisted of multiple test forms. Each examinee responded to one random test form and received a test booklet consisting of two blocks of items: a block that was included in all booklets, and a matrix-sampled block that was specific only to one form. Each block included both dichotomously (*0* or *1*) and polytomously (*0* to *k, k*>*1*) scored items. In a typical data matrix for a test for one year, examinee responses were arranged in rows and item responses in columns as shown in Figure 1.1. For example, examinees taking Form 3 took the common block and the matrix-sampled Block 3, represented by the two shaded cells. The common block should not be confused with the common/equating items. The common block of items across the forms in one year consists of items that adhere to the test specifications on content coverage and, at the same time, serves the purpose of placing all forms on a single scale on a simultaneous calibration of all forms. The common items used for equating, the common-item pool that appeared in the previous year administration, were arranged in the matrix-sampled blocks.



Common block	Matrix-sampled blocks of items								Form
	1	2	3	4	5	6	7	8	
■	■								Form 1
■		■							Form 2
■			■						Form 3
■				■					Form 4
■					■				Form 5
■						■			Form 6
■							■		Form 7
■								■	Form 8
	Common/equating items are arranged in the matrix-sampled blocks								

Figure 1.1. Matrix-sampled design for an eight-form annual administration.

Both dichotomous and polytomous items comprised the common-item pool as well. Since they were arranged in the matrix-sampled part of each form, each examinee encountered only a subset of the common items. The number of forms, the total number of items and the number of common items in each assessment appear on Table 1.2.

Table 1.2

Item Information for the Four Assessment Programs

Assessment	Number of forms (Year 1/Year 2)	Total number of items (Year 1/Year 2)	Number of common items between Year 1 and Year 2
Mathematics 8	8/8	139/137	44
Science 11	12/12	126/138	45
Social Studies 6	8/12	124/95	56
Science 6	8/12	123/95	50

Methodology

Overview

The delta-plot method uses item difficulties of the common items to identify the items with unexpected change in their difficulty across administrations. The Year-1 and the Year-2 administrations are calibrated separately to obtain item and ability parameter estimates. The estimates of the common items are used to derive an equating transformation that will rescale the Year-2 scale to the Year-1 scale. A second transformation is derived when the common item parameters of the outliers are discarded. The effects of the outliers on the equating of Year-2 score distribution will be examined by looking at the average of the equated distributions and the percentage of students above certain cut-points under the two common-item sets. This procedure is followed for each of the four assessments and applying (a) the Stocking and Lord, (b) the Haebara, (c) the mean/sigma, and (d) the mean/mean IRT equating methods. A 3PL IRT model for the dichotomous items and a graded response model for the polytomous items is used to calibrate the data. In addition, a 1PL model is fitted and equating is carried out with the same equating methods. In total, there are four assessments and eight model-by-equating method combinations. The procedures are described in detail below.

The Delta-Plot Method

In the delta-plot procedure, p_{Yj}, the proportion correct of a common item j in administration Y (here Y=$1,2$) is transformed to the delta metric through a linear transformation of the inverse normal equivalent (Dorans and Holland, 1993).

$$\delta_{Yj} = 13 - 4\,\Phi^{-1}(p_{Yj}) \qquad (1.1)$$

In the delta metric, a p-value of 0.5 is transformed to 13, larger delta values correspond to more difficult items and smaller delta values to easier items, as opposed to the proportion correct scale, which is bounded between 0 and 1 with easier items having higher values than more difficult ones.

Two groups respond to the same items, the item p-values, p_{Yj}, for each group are calculated, transformed to the delta metric with Equation 1.1, and plotted on a scatterplot. Each point corresponds to an item with a delta value, $_{1j}$, for the Year-1 group plotted on the horizontal axis and a delta value, $_{2j}$, for the Year-2 group plotted on the vertical axis. Outliers denote items that are functioning differentially for the two groups with respect to the level of difficulty.

A handy rule to determine which items are outliers is by drawing a "best-fit" line to the points and calculating the perpendicular distances of each point to the line. The fitted line is chosen to minimize the sum of squared perpendicular distances (not the sum of squared vertical distances as in ordinary least squares regression) of the points to the line.[5] Any point lying more than three standard deviations of the distances away from the line is a candidate for exclusion from the common item pool. Such items call for inspection to determine plausible causes for the differential performance in the two groups.

To carry out the delta-plot procedure the Year-1 and Year-2 p-values for each common item were calculated. For the dichotomous items, the p-value is the proportion of correct responses; for the polytomous items, it is the mean score that all examinees obtained on the item divided by the maximum possible score. For each pair of administrations, e.g., the Mathematics Grade 8 in Year 1 and Year 2, a delta-plot was constructed graphing the Year-1 versus the Year-2 p-values transformed to the delta

[5] This is know as "principal components" or "principal axis" regression and unlike ordinary least squares regression, it is symmetric: the line obtained by regressing the independent on the dependent variable and the line obtained by regressing the dependent on the independent variable are identical.

metric with Equation 1.1. A straight line was fitted in each plot to identify outlying points as shown in Equation 1.2. The slope for the best-fit line was estimated by the ratio of standard deviations of the δ_{Yj}, and the intercept was determined such that the line passes through the point where the abscissa is the mean of the transformed p-values for Year 1 and the ordinate is the mean of the transformed p-values for Year 2.

$$y = \frac{s(\delta_{2j})}{s(\delta_{1j})} x + \bar{\delta}_{2j} - \frac{s(\delta_{2j})}{s(\delta_{1j})} \bar{\delta}_{1j} \qquad (1.2)$$

The distance of each point to the best-fit line was then calculated. Any points lying more than three standard deviations of all distances away from that line were flagged as outliers. Thus, two sets of items are defined: (a) all common items, and (b) the non-outlying common items. If the equating transformations based on the two sets differ markedly, then the inclusion or exclusion of outliers from the common-item pool could be influential on the results of equating.

Test Calibrations and IRT Models

Prior to estimating equating transformations, each assessment was calibrated separately with PARSCALE 3.0 (Muraki & Bock, 1997). The assessments came in multiple forms every year with certain items embedded in all those forms – in addition to the common items embedded across years for equating cohort scores. Concurrent calibration of all forms for a single year automatically places them on one scale.

PARSCALE 3.0 allows calibration of tests that consist of both dichotomous and polytomous items. First, a 3PL IRT model (Birnbaum, 1968) was fitted to the dichotomous data as in Equation 1.3.

$$P_j(\theta_i) = c_j + (1 - c_j) \frac{e^{Da_j(\theta_i - b_j)}}{1 + e^{Da_j(\theta_i - b_j)}} \qquad (1.3)$$

P$_j$($_i$) is the probability that an examinee *i* with ability level $_i$ responds to item *j* correctly. *D* is a constant equal to 1.7. For each item *j*, the model provides estimates for the three parameters a_j, b_j, and c_j; the discrimination, difficulty or location, and pseudoguessing or lower asymptote parameters, respectively. It also provides an estimate of the examinee's ability level $_i$. The dichotomous and the polytomous items were all scaled together. For the polytomous items, a graded response model (Samejima, 1969) was fitted. The graded response model applies to items that are scored in ordered categories with higher categories representing better performance than lower categories. The probability of an examinee scoring in a category *k (k=0,…,m)* is equal to the probability of obtaining a score of *k* or above, $P_{jk}^{+}(\theta)$, minus the probability of obtaining a score of *k+1* or above, $P_{j,k+1}^{+}(\theta)$. Its logistic form given by Muraki and Bock (1997) is

$$P_{jk}(\theta) = P_{jk}^{+}(\theta) - P_{j,k+1}^{+}(\theta) = \frac{e^{Da_j(\theta - b_j + d_k)}}{1 + e^{Da_j(\theta - b_j + d_k)}} - \frac{e^{Da_j(\theta - b_j + d_{k+1})}}{1 + e^{Da_j(\theta - b_j + d_{k+1})}} \qquad (1.4)$$

where $P_{jk}(\ )$ is the probability that an examinee with ability level obtains a score *k* on an item *j* with *m+1* scoring categories; d_k is the category parameter, with $\sum_{k=0}^{m} d_k = 0$, and $d_k - d_{k+1} \geq 0$. The difference b_j-d_k is referred to as the *category-threshold* parameter, the location on the scale that separates two adjacent scores. The probability of responding in one of the two extreme categories or above is defined as $P_{j0}^{+}(\theta) = 1$ and $P_{j,m+1}^{+}(\theta) = 0$. A polytomous item has *m* category-threshold parameters (separating the *m+1* scoring categories), as opposed to a single location parameter for a dichotomous item. When a polytomous item appears in a common-item pool, it will contribute all its category-threshold parameters, and will thus have more weight in the IRT equating transformation than the dichotomous items. This is not undesirable, since a polytomous item carries more weight, i.e., may contribute more score points, in the estimation of an individual examinee's score.

A 1PL model was also fitted,

$$P_j(\theta_i) = \frac{e^{D(\theta_i \text{-} b_j)}}{1 + e^{D(\theta_i \text{-} b_j)}}$$

by fixing the a_j parameters at 1 and excluding the c_j from the model. The graded response model (Equation 1.4 with the a_j parameters fixed at 1) was fitted to the polytomous data.

PARSCALE 3.0 estimates the parameters of the models by marginal maximum likelihood and the EM algorithm is used in the solution of the likelihood equations

(Muraki & Bock, 1997). Often the calibrations in PARSCALE 3.0 did not converge. Some items that had large standard errors in their IRT location parameters, or for which very few examinees had selected one of the responses, were dropped from the calibration to enable the convergence of the procedure. The parameters of the dropped items were fixed to the initial values and were not estimated after each EM cycle. In the data sets presented in this study none of the dropped items was a common item.[6]

Equating Methods

When two tests with embedded common items are independently calibrated, and their ability and difficulty parameter scales arbitrarily fixed, their scales are not identical, but linearly related (Lord, 1980) and a linear transformation of the latent scale does not affect the predicted probability of correct response to an item. If *I* and *J* are the two scales (or groups), then the ability and IRT b_j parameters of the two scales are related linearly

$$\theta_{Ji} = A\theta_{Ii} + B, \tag{1.5}$$

$$b_{Jj} = Ab_{Ij} + B, \tag{1.6}$$

for individuals *i* and items *j*. Assuming a 3PL IRT model, the a_j and c_j parameters are related as follows:

$$a_{Jj} = \frac{a_{Ij}}{A}, \tag{1.7}$$

$$c_{Jj} = c_{Ij} \tag{1.8}$$

Moments methods make use of the means and standard deviations of the common item parameters to estimate the constants A and B. The mean/sigma method (Marco, 1977) uses the moments of the b_j parameter estimates in place of the parameters below:

[6] More than the four assessments presented here were available for analysis, but their calibrations were more problematic: either they did not converge, or when they did, some of the common items had to be skipped, which implies that the parameter estimates of the skipped items were not estimated but retained their initial values. Equating transformations are derived based on common-item parameter estimates. Badly estimated parameters would influence equating in extraneous ways. The four data sets in this study can be considered as relatively "well-behaved" in this respect. The problematic data sets were not analyzed further.

$$A = \frac{\sigma(b_{Jj})}{\sigma(b_{Ij})},$$

$$B = \mu(b_{Jj}) - A\mu(b_{Ij})$$

The mean/mean method (Loyd & Hoover, 1980) estimates the intercept *B* in the same way, but the slope *A* as the ratio of the means of the IRT a_j parameter estimates

$$A = \frac{\mu(a_{Ij})}{\mu(a_{Jj})}$$

Characteristic curve transformation methods consider information from all item parameters simultaneously by minimizing the difference between item characteristic curves. From Equations 1.5 to 1.8, two scales *J* and *I* are related by

$$P_{ij}(\theta_{Ji}; a_{Jj}, b_{Jj}, c_{Jj}) = P_{ij}(A\theta_{Ii} + B; \frac{a_{Ij}}{A}, Ab_{Ij} + B, c_{Ij})$$

for examinee *i* and item *j*. When the estimates of *A* and *B* are used, the above equality does not hold exactly for all items and examinees (Kolen & Brennan, 2004). Characteristic curve transformation methods estimate the *A* and *B* such that the difference between the left and the right part of the equation is minimum. The method developed by Haebara (1980) minimizes the sum over all examinees *i* of the sum of squared differences between item characteristic curves over all the common items *j*

$$\sum_i \left\{ \sum_j \left[P_{ij}(\theta_{Ji}; \hat{a}_{Jj}, \hat{b}_{Jj}, \hat{c}_{Jj}) - P_{ij}(\theta_{Ji}; \frac{\hat{a}_{Ij}}{A}, A\hat{b}_{Ij} + B, \hat{c}_{Ij}) \right]^2 \right\}$$

The Stocking and Lord (1983) procedure chooses the constants A and B such that the function

$$F = \frac{1}{N} \sum_i \left[\sum_j P_{ij}(\theta_{Ji}; \hat{a}_{Jj}, \hat{b}_{Jj}, \hat{c}_{Jj}) - \sum_j P_{ij}(\theta_{Ji}; \frac{\hat{a}_{Ij}}{A}, A\hat{b}_{Ij} + B, \hat{c}_{Ij}) \right]^2$$

is minimized over *i* examinees and *j* common items

Stocking & Lord, Haebara, mean/sigma, and mean/mean were the four equating methods used in this study. The first two methods are characteristic curve transformation methods and estimate the equating coefficients to minimize the area between item characteristic curves, weighted according to the examinee ability distribution. Under a 3PL IRT model, all three parameter estimates of the common items are taken into account.[7] The 1PL IRT model does not allow for guessing, and constrains the discrimination parameters to unity; the item characteristic curves differ only with respect to their location.

The latter two methods are based on the moments of the item parameters. The mean/sigma depends only on the difficulty parameter estimates, while the mean/mean incorporates information from the both the difficulty and the discrimination parameters. The mean/mean method may not be appropriate with a 1PL IRT calibration since discrimination parameters are fixed to one for all times.

Different combinations of item parameters in the 3PL IRT model can result in similar item characteristic curves (Hulin, Lissak, & Drasgow, 1982). Given an empirical item characteristic curve, there is an interaction in the estimates of the three item parameters. The variety of item parameters that can be described by a given item characteristic curve is more pronounced for easy and hard items, because there are less data available in large sections of the ability scale to define the item characteristic curve. IRT alignment procedures that rely on individual item parameters, such as the mean/sigma and the mean/mean equating methods, as opposed to the more informative item characteristic curves alignment methods, are more sensitive to trade-offs in the item parameter estimates. To improve the stability of the discrimination and the difficulty parameters (a_j, b_j) estimates on which moments equating methods rely on, the lower asymptote (c_j) estimates of the common items derived from the Year-1 calibration were held fixed in the Year-2 calibration (W. Yen, personal communication, June 24 2003).

The equating transformations were estimated using the ST software (Hanson & Zeng, 1995), which takes the common-item parameter and the ability distribution estimates from two administrations as input. The output returns estimates for the slope and intercept of the linear transformations (see Appendix 1) In the 3PL cases, each common item contributed three parameter estimates: discrimination, difficulty and

[7] As will be described shortly, because the lower asymptote parameters were constrained in the Year-2 calibration to their Year-1 values to improve the software's estimation procedures, it is the discrimination and difficulty parameter estimates of the common items that differ across years.

lower asymptote. In the 1PL case, all common item discriminations were equal to one and all lower asymptotes to zero. Polytomous items contributed more than just one set of parameters. Those items were typically scored from zero to four, so they had four category-threshold estimates. In the ST input file one such item would have four entries, one for each category-threshold estimate. The discrimination parameter estimate was entered four times as an identical value, while the lower asymptote was kept to zero.

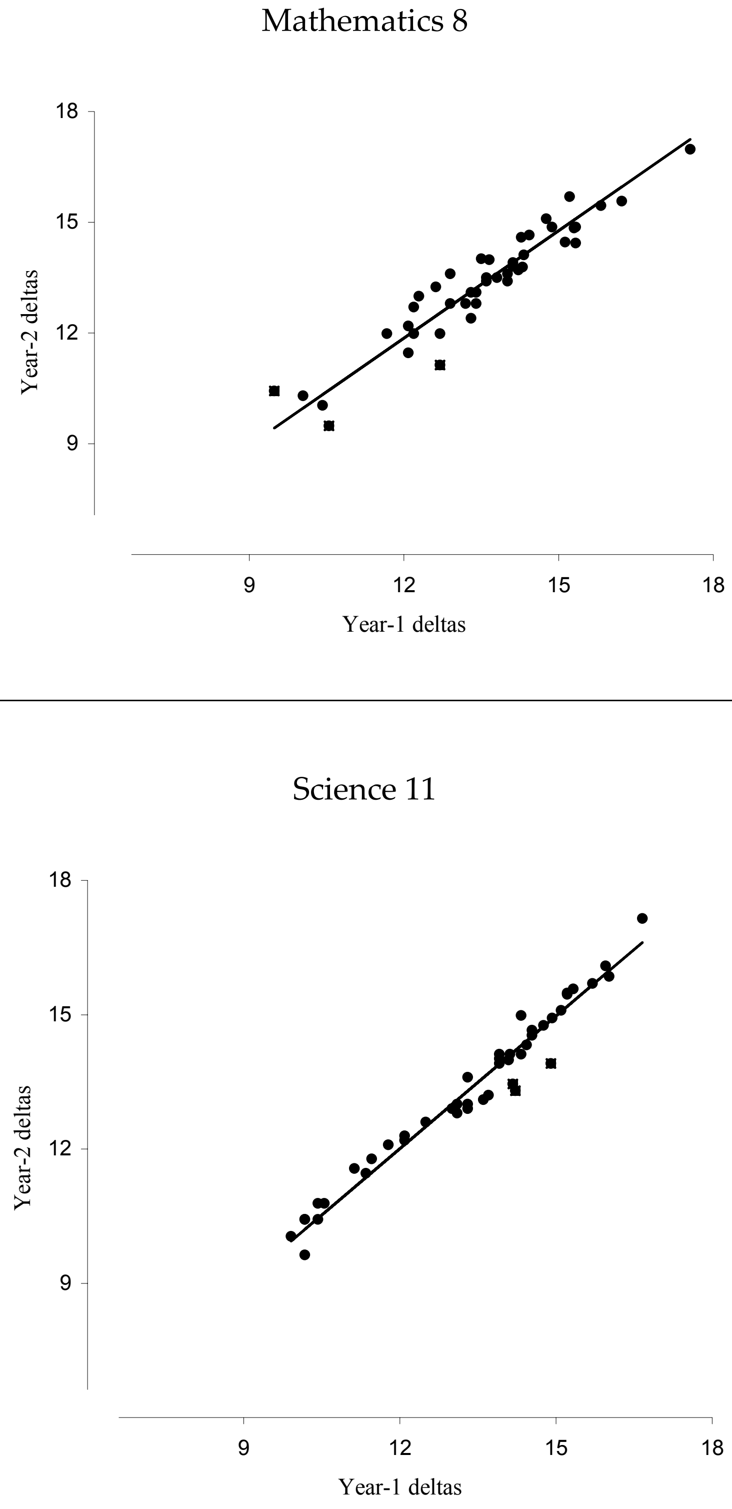
Results

Delta-Plots

Figure 1.2 displays the delta-plots for each of the four assessments. Each graph depicts the Year-1 and Year-2 delta values for the common items, the best-fit line, and the outliers identified with square markers. All the outliers were dichotomous items with the exception of the two rightmost outlying items on the Science 11 assessment. Flagged points that appear below the line represent items that became unexpectedly easier in the second administration of the assessment, while those that appear above the line stand for items that increased in their difficulty from Year 1 to Year 2.

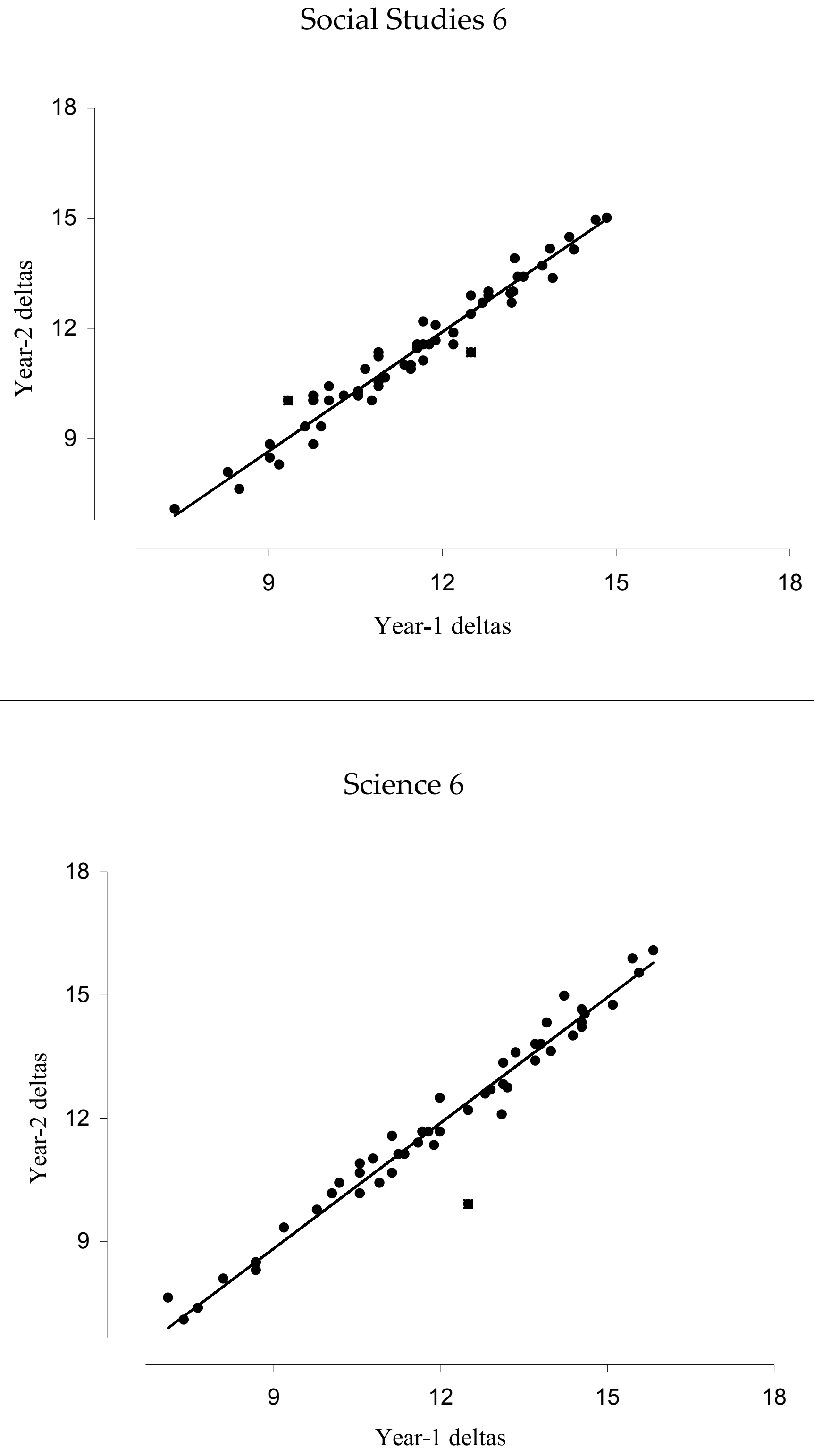
Ability Distributions Before Equating

The theoretical distribution of the ability parameters is normal with mean zero and standard deviation one. The moments of the estimated examinee ability distributions for each administration, prior to equating, are presented in Table 1.3. The upper panel tabulates the moments for the distributions obtained from a 3PL IRT model fit. All ability distributions have estimated means that are just below zero, and standard deviations that are less than one. They are all negatively skewed and slightly more peaked than the normal distribution. The moments of the ability distributions estimated by a 1PL IRT model appear on the lower panel of Table 1.3. The eight distributions are centered closer to zero than those under the 3PL model, but have much smaller standard deviations. Hence, they have large kurtosis values; they are very peaked and most of them are negatively skewed.


Mathematics 8

Science 11

Figure 1.2. Delta-plots of the common items for the four assessments.



Social Studies 6

Science 6

Figure 1.2. Delta-plots of the common items for the four assessments (continued).

Table 1.3

Moments of the Ability Distributions

Assessment	Year	Mean	Standard deviation	Skewness	Kurtosis	N
Calibration with a 3PL IRT model						
Mathematics 8	2	-0.002	0.957	-0.259	0.169	7128
	1	-0.121	0.855	-0.397	0.425	7258
Science 11	2	-0.002	0.931	-0.111	-0.110	14565
	1	-0.126	0.930	-0.436	0.451	14244
Social studies 6	2	-0.187	0.944	-0.301	0.469	17371
	1	-0.047	0.948	-0.370	0.361	17126
Science 6	2	-0.107	0.908	-0.436	0.637	17371
	1	-0.033	0.923	-0.360	0.469	17128
Calibration with a 1PL IRT model						
Mathematics 8	2	0.014	0.665	0.016	0.616	7128
	1	0.018	0.704	-0.198	1.284	7258
Science 11	2	-0.010	0.536	0.121	0.247	14565
	1	-0.008	0.584	-1.008	3.610	14244
Social studies 6	2	0.006	0.542	-0.606	3.131	17371
	1	0.010	0.587	-0.601	2.214	17126
Science 6	2	0.009	0.577	-0.663	2.657	17371
	1	0.007	0.568	-0.546	2.294	17128

The spread and location of the IRT scale are arbitrarily determined in each calibration. The Year-1 and Year-2 scores of an assessment are not comparable when they are calibrated separately. If an appropriate equating transformation is applied on the Year-2 scale, the scores on the two scales will be interchangeable. When a scaling transformation based on information from the common items is applied on the Year-2 scale, the scores on the two scales will be interchangeable.

For each of the four assessments, two models are fitted (3PL and 1PL), resulting in eight pairs (Year-1 and Year-2) of distributions. Equating of the Year-2 scores to the Year-1 scale is performed either using all the common items to estimate the slope and intercept of the linear transformations or using only the non-outlying delta-plot common items. The transformations are calculated with four different methods. Appendix 1 presents the ST output files with the slopes and intercepts of the equating transformations.

Effects of the Common-Item Set Used on the Year-2 Means and Annual Gains

A series of four tables that follow illustrate the effect of equating using all the common items, versus excluding the delta-plot outlying items, on the Year-2 equated mean score. The effect is examined under eight conditions: two types of IRT models by four IRT equating methods. Once the Year-2 mean score is scaled to the anchor scale by a linear transformation ($\bar{\hat{\theta}}_2^* = A\bar{\hat{\theta}}_2 + B$), it can be directly compared with the previous year mean, $\bar{\hat{\theta}}_1$. Gain (or decline) is the difference of the two means. The standard error of this quantity, treating the coefficients *A* and *B* as fixed, is calculated as shown below:

$$SE\left(\bar{\hat{\theta}}_2^* - \bar{\hat{\theta}}_1\right) = \sqrt{Var\left(\bar{\hat{\theta}}_2^* - \bar{\hat{\theta}}_1\right)} = \sqrt{Var\left(\bar{\hat{\theta}}_2^*\right) + Var\left(\bar{\hat{\theta}}_1\right)} =$$

$$\sqrt{Var\left(A\bar{\hat{\theta}}_2 + B\right) + Var\left(\bar{\hat{\theta}}_1\right)} = \sqrt{A^2 Var\left(\bar{\hat{\theta}}_2\right) + Var\left(\bar{\hat{\theta}}_1\right)} =$$

$$\sqrt{\frac{A^2}{N_2^2} Var\left(\sum_{i=1}^{N_2} \hat{\theta}_{2i}\right) + \frac{1}{N_1^2} Var\left(\sum_{i=1}^{N_1} \hat{\theta}_{1i}\right)} =$$

$$\sqrt{\frac{A^2}{N_2^2} \sum_{i=1}^{N_2} Var\left(\hat{\theta}_{2i}\right) + \frac{1}{N_1^2} \sum_{i=1}^{N_1} Var\left(\hat{\theta}_{1i}\right)} =$$

$$\sqrt{\frac{A^2}{N_2} Var\left(\hat{\theta}_2\right) + \frac{1}{N_1} Var\left(\hat{\theta}_1\right)}$$

The tables finally show a ratio that compares the gain under equating with all common items to the gain under equating with the non-outliers, taking the latter as the baseline.

Table 1.4(a) displays results for the Mathematics Grade 8 assessment. There were three outliers on the delta-plot of this assessment, all in the lower end of the difficulty scale: two common items were much easier and one was more difficult for the Year-2 cohort. When these outliers are included in the equating, the Year-2 mean is higher than when they are not; the Year-2 cohort shows better performance overall when the outliers are included and hence the estimated transformations (under any method or model) favors them. Similarly the gain from Year 1 to Year 2 is slightly larger (or the decline is smaller) when the outliers are included in the equating. There is a discrepancy between the two fitted models. The 1PL model shows positive gains at the magnitude of twice the standard error of the gain under all four methods consistently. The use or not of the outliers affects gains in the expected direction but only minimally as can be seen from the ratio of gains under the two sets of common items. In contrast, results are not consistent when a 3PL IRT model is fitted. There is a very small decline when the Stocking & Lord method is used, but a larger one with the Haebara and the mean/mean methods and a similar in magnitude gain with the mean/sigma method.

Table 1.4(a)

Means and Gains for Mathematics 8

IRT model	3PL				1PL			
Equating method	S-L[1]	H[1]	M/M[1]	M/S[1]	S-L	H	M/M	M/S
Year-2 mean (all items)	-0.1221	-0.1364	-0.1331	-0.1049	0.0431	0.0430	0.0383	0.0359
Year-2 mean (non-outliers)	-0.1294	-0.1449	-0.1340	-0.1108	0.0428	0.0427	0.0379	0.0353
Gain/decline[2] (all items)	-0.0013	-0.0155	-0.0123	0.0159	0.0247	0.0247	0.0200	0.0176
SE	0.0142	0.0143	0.0142	0.0139	0.0115	0.0116	0.0114	0.0115
Gain/decline (non-outliers)	-0.0085	-0.0241	-0.0131	0.0100	0.0244	0.0243	0.0195	0.0170
SE	0.0142	0.0143	0.0141	0.0139	0.0115	0.0116	0.0114	0.0115
Ratio of gains	0.15	0.65	0.93	1.59	1.01	1.01	1.02	1.04

[1]S-L: Stocking & Lord, H: Haebara, M/M: mean/mean, M/S: mean/sigma

[2]Year 1 means appear on Table 1.3

Results for the Science 11 assessment appear on Table 1.4(b). There were three delta-plot outliers below the best-fit line, i.e., decreased in difficulty for the Year-2 examinees. Two of them were polytomous items; polytomous items weigh more in the equating since they contribute multiple category-threshold parameters (in this case, four each.) As expected, retaining the outliers in the common-item pool results in higher Year-2 means than when only non-outlying items are used. Irrespective of which method or model is applied, there are large gains in Year 2, but when the outliers are discarded, the gains are much smaller and in some cases result in minor declines.

Table 1.4(b)

Means and Gains for Science 11

IRT model	3PL				1PL			
Equating method	S-L	H	M/M	M/S	S-L	H	M/M	M/S
Year-2 mean (all items)	-0.0874	-0.0751	-0.0991	-0.0744	0.0379	0.0330	0.0424	0.0418
Year-2 mean (non-outliers)	-0.1493	-0.1412	-0.1754	-0.1200	0.0018	0.0049	-0.0114	0.0020
Gain/decline (all items)	0.0384	0.0508	0.0267	0.0514	0.0457	0.0408	0.0502	0.0496
SE	0.0107	0.0107	0.0106	0.0104	0.0066	0.0067	0.0066	0.0066
Gain/decline (non-outliers)	-0.0235	-0.0154	-0.0496	0.0058	0.0096	0.0127	-0.0036	0.0098
SE	0.0108	0.0107	0.0108	0.0103	0.0066	0.0066	0.0066	0.0065
Ratio of gains	-1.64	-3.30	-0.54	8.88	4.74	3.21	-14.04	5.04

The delta-plot for the Social Studies 6 assessment showed two outliers, both dichotomous items; one was above the best-fit line and one below. As shown on Table

1.4(c), the Year-2 mean is very similar whether the outliers are kept in the common-item pool or not. All four equating methods under both IRT models show a consistent decline from Year 1 to Year 2 that ranges from 2 to 3.5 times the standard error of the gain. The ratios of gains demonstrate that the influence of the decision on how to treat the delta-plot outliers is negligible.

In Science Grade 6 there was just one delta-plot outlying item of moderate difficulty on which the Year-2 cohort performed strikingly better then their Year-1 counterparts. The Year-2 mean score is higher if that outlier is retained in the equating than when it is not. The gain is also larger (or the decline smaller.) As can be seen in Table 1.4(d), with a 1PL calibration the gains are consistent across methods. With a 3 PL calibration the Haebara method shows a gain, while the other methods and particularly the moments methods show a decline from Year 1 to Year 2.

Table 1.4(c)

Means and Gains for Social Studies 6

IRT model	3PL				1PL			
Equating method	S-L	H	M/M	M/S	S-L	H	M/M	M/S
Year-2 mean (all items)	-0.0730	-0.0706	-0.0770	-0.0711	-0.0088	-0.0057	-0.0101	-0.0104
Year-2 mean (non-outliers)	-0.0730	-0.0702	-0.0772	-0.0710	-0.0093	-0.0061	-0.0105	-0.0108
Gain/decline (all items)	-0.0260	-0.0236	-0.0300	-0.0241	-0.0190	-0.0159	-0.0203	-0.0206
SE	0.0101	0.0101	0.0101	0.0100	0.0061	0.0061	0.0061	0.0061
Gain/decline (non-outliers)	-0.0260	-0.0232	-0.0302	-0.0240	-0.0195	-0.0163	-0.0206	-0.0210
SE	0.0101	0.0101	0.0101	0.0100	0.0061	0.0061	0.0061	0.0061
Ratio of gains	1.00	1.02	0.99	1.00	0.98	0.97	0.98	0.98

Table 1.4(d)

Means and Gains for Science 6

IRT model	3PL				1PL			
Equating method	S-L	H	M/M	M/S	S-L	H	M/M	M/S
Year-2 mean (all items)	-0.0511	-0.0166	-0.0792	-0.0752	0.0173	0.0140	0.0211	0.0224
Year-2 mean (non-outliers)	-0.0559	-0.0204	-0.0838	-0.0800	0.0159	0.0123	0.0200	0.0213
Gain/decline (all items)	-0.0177	0.0169	-0.0458	-0.0418	0.0107	0.0074	0.0145	0.0158
SE	0.0095	0.0096	0.0095	0.0094	0.0062	0.0063	0.0062	0.0061
Gain/decline (non-outliers)	-0.0225	0.0130	-0.0504	-0.0466	0.0093	0.0058	0.0134	0.0147
SE	0.0095	0.0096	0.0095	0.0094	0.0062	0.0063	0.0062	0.0061
Ratio of gains	0.79	1.30	0.91	0.90	1.16	1.29	1.08	1.07

Effects of the Common-Item Set Used on the Percent Above a Cut-Off

The percent above a cut-off is a statistic that educators often refer to in the context of testing. For the purposes of standards-based reporting, individuals are assigned to performance categories based on their achievement on an assessment program. These performance categories have labels such as “below basic,” “basic,” “proficient,” and “advanced” associated with them, as well as performance descriptions of what a student is expected to be capable of doing within each category. The cut-off points on the achievement scale that serve as borderlines between two adjacent categories are determined through standard setting procedures. The percentage of students in a category or above has become a popular descriptor of how well a school, a district, or a state is doing. In addition to more traditional ways of score reporting (averages and longitudinal gains for a population and its subpopulations), progress is tracked through the trends in the percentage of students at, above, or below a category. Also, targets are set with respect to how high a percentage should be reached within a time period.

Cut-off points between categories are not fixed; they are specific to individual assessments. For illustration purposes, two points on the theta scale *0* and *0.5* will serve as arbitrary cut-offs to examine how the percentage varies in the Year-2 ability distribution depending on whether the delta-plot outliers are kept in the common-item pool or not. The percentage is empirically determined by counting the number of examinees with an estimate of ability (after equating) above the cut-off point and dividing by the total number of examinees. For ease of presentation, only two of the four equating methods, the Stocking & Lord and the mean/sigma, are presented.

In the Mathematics Grade 8 assessment under a 3PL IRT model, the percents above a cut-off decrease in Year 2 compared to Year 1 (Table 1.5(a)). The decrease is larger when the outliers are discarded from the common-item pool. In the previous section, Table 1.4(a), inclusion of the common items resulted in larger gains or smaller decreases in the mean score. But the mean/sigma method showed a small increase in the mean, while the percent above $\hat{\theta}=0$ goes down from Year 1 to Year 2 by more than 1%. This might be due to the different shapes of the two ability distributions. The Year-1 distribution is more leptokurtic and negatively skewed than the Year-2 distribution. With a 1PL calibration, the percents above a cut-off increase in Year 2 compared to Year 1; they are slightly higher when outliers are kept in the equating.

Table 1.5(a)

Percentage Above Cut-Off for Mathematics 8

	IRT model	3PL		1PL	
	Equating method	S-L	M/S	S-L	M/S
Percent above $\hat{\theta}$=0	Year 1	47.00%		50.84%	
	Year 2 (all items)	45.37%	45.88%	51.89%	51.40%
	Year 2 (non-outliers)	44.88%	45.62%	51.87%	51.36%
Percent above $\hat{\theta}$=0.5	Year 1	23.23%		23.22%	
	Year 2 (all items)	23.33%	22.94%	23.99%	23.53%
	Year 2 (non-outliers)	22.95%	22.71%	23.99%	23.51%

Results for Science Grade 11 on Table 1.5(b), as in the case of mean scores, show a large influence of the outlying common items on the percent above a cut-off. Across models and methods, inclusion of the outlying common items seems to benefit 2-3% of the population by re-classifying them above instead of below the cut-off.

Table 1.5(b)

Percentage Above Cut-Off for Science 11

	IRT model	3PL		1PL	
	Equating method	S-L	M/S	S-L	M/S
Percent above $\hat{\theta}$=0	Year 1	46.70%		51.68%	
	Year 2 (all items)	46.75%	47.11%	51.76%	52.10%
	Year 2 (non-outliers)	44.02%	44.74%	48.99%	49.04%
Percent above $\hat{\theta}$=0.5	Year 1	25.26%		16.46%	
	Year 2 (all items)	26.29%	25.12%	18.87%	18.99%
	Year 2 (non-outliers)	24.38%	22.71%	16.92%	15.92%

Table 1.5(c) tells the same story as Table 1.4(c) with a minimal influence of the outliers on the percent above a cut-off. All model-by-method results are consistent in showing a decrease in the Year-2 percents.

Table 1.5(c).

Percentage Above Cut-off for Social Studies 6

	IRT model	3PL		1PL	
	Equating method	S-L	M/S	S-L	M/S
Percent above $\hat{\theta}$=0	Year 1	50.30%		52.60%	
	Year 2 (all items)	47.99%	48.03%	49.87%	49.76%
	Year 2 (non-outliers)	47.99%	48.03%	49.85%	49.73%
Percent above $\hat{\theta}$=0.5	Year 1	28.98%		18.87%	
	Year 2 (all items)	26.99%	26.83%	15.52%	15.48%
	Year 2 (non-outliers)	27.01%	26.83%	15.49%	15.47%

For Science 6, keeping the outlying common items has a small positive influence on the percent above $\hat{\theta}$*=0* and $\hat{\theta}$*=0.5* of less then 0.3%. When a 3PL model was fitted the percents declined in Year 2, as opposed to an increase when a 1PL model was fitted. The patterns on Table 1.5(d) are consistent with those on Table 1.4(d).

Discussion

This study explored the effects of common item selection on aggregate score results. Two clusters of common items were considered in each of four statewide assessments: use of all common items to produce the equating transformation and use of those that do not indicate anomalous behavior on the delta-plot. Excluding items on which student cohorts perform too differentially compared to their performance on other items is an issue of face validity and fairness. Given that the delta-plot procedure is just a handy method to identify the few items that behave anomalously across cohorts of students, and that the “three standard deviations away from the line” rule is somewhat arbitrary, it might deserve more consideration whether the delta-plot outliers should be discarded from or kept in the common-item pool.

Table 1.5(d)

Percentage Above Cut-off for Science 6

	IRT model	3PL		1PL	
	Equating method	S-L	M/S	S-L	M/S
Percent above $\hat{\theta}=0$	Year 1	50.31%		51.86%	
	Year 2 (all items)	50.23%	48.89%	53.51%	53.87%
	Year 2 (non-outliers)	50.00%	48.62%	53.38%	53.75%
Percent above $\hat{\theta}=0.5$	Year 1	28.67%		17.83%	
	Year 2 (all items)	25.85%	23.97%	18.55%	18.34%
	Year 2 (non-outliers)	25.58%	23.67%	18.50%	18.28%

The analysis presented in this study relied on real data. There were no controlled simulated conditions that would demonstrate which characteristics of the common items are influential on summary statistics such as the mean, the mean gain across years, and the percent above a cut-off point. However, these four actual cases suggest a number of possible characteristics that would help predict whether exclusion of items from the common-item pool would be influential or not. These characteristics include the number of outlying common items, their type (dichotomous or polytomous), whether the outliers lie above, below, or on both sides of the best-fit line, the level of difficulty of the outliers, the IRT model used for calibration of the data, and the equating method.

Very few common items, less than 7% of the common-item pool in all the cases presented above, were identified as outliers. The number of the outliers by itself though is not very informative as to what the influence might be on equated summary statistics. In the case of Science 11 two of the three outliers were polytomous and each one of those contributed four sets of parameter estimates in the equating transformation. If there are polytomous items in the common-item pool which behave differentially across administrations, discarding them from the equating process will likely have a considerable impact on aggregates, more so than dichotomous items, as was seen in the Science 11 assessment results.

Inclusion of outlying common items benefits the group that performs better on them. In the Science 11 assessment all three outliers were below the scatter of the other items; this meant that the Year-2 cohort performed substantially better on them than did the Year-1 cohort, so including them in the common-item pool gives an advantage to the Year-2 cohort. In contrast, the two outliers in the Social Studies Grade 6 assessment lay on both sides of the best-fit line. It is likely that in this case their effects canceled out and as a result, the summary statistics did not vary much with regard to the treatment of the outliers.

The location of an outlier on the difficulty scale could also determine how influential it is; this idea is similar to the notion of leverage in regression analysis. If an item is very easy or very difficult and is also flagged as an outlier, then it will probably be more influential than an outlier of moderate difficulty. This conjecture is supported by the comparison between the one assessment that demonstrated large effects on summary statistics and the two that did not. The Science 11 involved three outliers that were above the average difficulty. In contrast, the Social Studies 6 involved two outliers one of which was relatively easy and the other relatively difficult, and the Science 6 test just one outlier, which although it was very distant from the rest of the items and thus would be expected to exert high influence on the transformation, lay at the middle of the difficulty scale and had a rather small impact on the aggregates. The Mathematics 8 test involved three outliers of low difficulty, but one was above the line and the other two below and this may have offset their influence.

In the Science 11 assessment where the treatment of the outliers had an obvious effect on the summary statistics, as well as in the Social Studies 6 test where their influence was small, all equating methods and both the 3PL and 1PL calibrations tended to show consistent outcomes. In the other two assessments the methods under the 1PL calibration were consistent, but when a 3PL model was fitted to the data, the different methods did not always produce similar results. A plausible explanation might relate to how accurately item parameters are estimated. With a 1PL IRT model, only item difficulties are estimated as opposed to the 3PL where more parameters need to be estimated. Thus when equating methods use information from more than one parameter, or the item characteristic curve, they maybe incorporating more information, however they are more sensitive to the uncertainties of the estimation of those parameters.

Some caveats are in order: the comments outlined above are merely conjectures that arise from the analysis and from knowledge of how outliers affect relationships. More formal analytical work, or simulations that control the above conditions, would be needed to establish more definite conclusions as to how delta-plot outliers affect equating. A second warning relates to the fact that the delta-plot outliers are determined from item p-values. Equating transformations are estimated from IRT parameter estimates or item characteristic curves. The p-values and the IRT *b* values are related – they are both difficulty indices. But the procedure carried out in this study, and which can be found in actual practice, involves two steps that function under two different frameworks. In theory, a delta plot outlier (based on p-values) might nonetheless satisfy the IRT parameter invariance assumption, whereas an item that looked fine in the delta plot might violate the invariance assumption.

The decision of how to deal with outliers has in some cases considerable effects on equating. In the study that follows an application of an alternative procedure for flagging outlying common items is presented. It is less crude, but still computationally easy to carry out, and enables more meaningful comparisons of performance on items, because item performance is examined for similar groups of examinees, not for overall student cohorts.

Study 2: The Mantel-Haenszel Procedure as an Alternative to the Delta-Plot Method

Overview

A straightforward and practical procedure used to flag items that do not behave similarly across administrations, groups, or years is the delta-plot method. It relies on item p-values and identifies items of inconsistent level of difficulty through outlier analysis. In this study a new method known in psychometrics for its applications to studies of DIF will be proposed for this purpose. The Mantel-Haenszel statistic (Mantel & Haenszel, 1959) estimates and tests the strength of association between two variables across levels of a third variable. In test analysis, it has been used to identify items that function differentially across gender, ethnicity, or other subpopulations; the association between performance on an item and membership in a subpopulation is studied across strata of examinees matched on some proxy of their proficiency.

Test administrations and their corresponding cohorts of students can be considered subpopulations and common items administered to more than one subpopulation can be examined for differential performance through Mantel-Haenszel procedures in a manner similar to the studies on DIF. For each common item, which will be referred to as the studied item, a *2x2* table can be constructed. The counts of correct and incorrect responses to the item in each one of two administrations/cohorts are tabulated in that table. The baseline administration group is called the reference group, and the other is called the focal group. A third variable, called the matching variable that is available for all subgroup members is used to match examinees on a measure of their proficiency. For each of the levels of the matching variable, there is a *2x2* table that contains counts for examinees of similar ability. Thus, the association between performance on the studied item and subgroup membership is examined for matched individuals.

Study 2 implements the Mantel-Haenszel procedure to flag common items that behave differentially across administrations. In the following sections, the data sources and the methodology used in Study 2 are first presented. Then, the results are reported and discussed. A comparison between the proposed and the delta-plot procedure explicates advantages and disadvantages of the new method.

Data Sources

Study 2 investigates the application of the Mantel-Haenszel procedure as an alternative to the delta-plot method to detect items that behave differentially across two administrations. A Visual and Performing Arts (VPA) Grade 4 statewide assessment is analyzed with both methods. Subjects covered include music, visual arts, dance, and theater. Students are assessed according to standards on creativity and expression, cultural heritage, and criticism and aesthetics.[8]

The test consisted of 12 forms and a total of 84 items. Each form comprised 7 items, of which 6 were multiple-choice dichotomous items, scored 0 or 1, and one was a polytomous, constructed-response question, scored on a scale of 0 to 4. The test was administered in both Year 1 and Year 2. Of the 84 items, 69 were common over the two annual administrations and distributed across the forms according to Table 2.1. The numbers of examinees taking each form are also presented on the same table.

[8] A non-typical assessment, such as VPA was selected for practical reasons discussed in the “Test design and implementation issues” section.

Methodology

Overview

First, the delta-plot procedure was applied to the test data and the outlying common items were identified, by plotting the p-values transformed to the delta metric and flagging outliers on the plot as described in study 1.

The Mantel-Haenszel procedure was applied next. The examinees taking Form X in Year 1 and those taking Form X in Year 2 constitute the reference and focal groups respectively. A number correct score for each examinee is derived by summing their scores on the common items. For example, the examinees taking Form 9 had number-correct scores ranging from 0 to 9, because there were 1 polytomous and 5 dichotomous common items in that particular form. The number-correct score serves as the matching criterion *j*.

Table 2.1

Characteristics of the 12 Forms of the VPA Grade 4 Assessment

Form	Total number of common items across years	Number of polytomous common items	Examinees taking the form in Year 1	Examinees taking the form in Year 2
1	5	1	1341	1335
2	6	1	1309	1316
3	7	1	1300	1316
4	6	1	1324	1300
5	5	0	1335	1312
6	7	1	1299	1297
7	5	1	1360	1323
8	6	-	1346	1308
9	6	1	1339	1311
10	5	1	1343	1313
11	6	1	1343	1315
12	5	0	1301	1296
Total	69	9	15940	15742

Treatment of the Dichotomous Items

For each dichotomous common item *i* in a form, a *2x2xK* three-dimensional table was constructed. One variable was the group each examinee belonged to (Year-1 or Year-2) and the second was his/her score on the dichotomous item *i* (0 or 1). The matching variable was the third dimension *j* of that table, *j=0,…, K*, where *K* is the maximum number-correct score on the common items of the form. As can be seen in Table Table 2.2, for one such partial table *j* for the dichotomous common item *i*, the counts of the correct (A_j or C_j) and incorrect (B_j or D_j) responses for each of the two examinee cohorts were recorded. For example, A_j would be the number of examinees in the Year-1 cohort with a number-correct score *j* on the common items of a form, who responded to item *i* correctly; B_j would be the number of examinees from that same group who got the item wrong.

Table 2.2

Counts for the j^{th} Partial Table for a Dichotomous Common Item *i*

		Score on item *i*		
		1	0	Total
Group	Year 1	A_j	B_j	N_{1j}
	Year 2	C_j	D_j	N_{2j}
	Total	M_{1j}	M_{0j}	T_j

The estimate of the Mantel-Haenszel common odds ratio for a common item *i* is then given by

$$\hat{\theta}_{MH} = \frac{\sum_j A_j D_j / T_j}{\sum_j B_j C_j / T_j} \qquad (2.1)$$

The common odds ratio takes values from 0 to infinity; with a $\theta_{MH} = 1$, there is no differential item performance between the two groups, and larger values imply that the item favors the reference group.

It is useful to refer to the natural logarithm of the common odds ratio, the "log-odds" for which there are approximate variance expressions. The log-odds ratio has a symmetric distribution centered at zero and for positive values of the log-odds the reference group performed better than matched examinees of the focal group; for negative values the opposite is true. Holland and Thayer (1988) report the following approximation of the variance of the log-odds derived by Phillips and Holland (1987)

$$Var\left[ln\left(\hat{\theta}_{MH}\right)\right] = \frac{1}{2\left(\sum_j A_j D_j / T_j\right)^2} \sum_j \left(A_j D_j + \hat{\theta}_{MH} B_j C_j\right)\left[A_j + D_j + \hat{\theta}_{MH}\left(B_j + C_j\right)\right] / T_j^{\ 2} \tag{2.2}$$

The quotient of the log-odds ratio with its standard deviation can be compared to the standard normal distribution for statistical significance. Both the magnitude and the statistical significance of the log-odds are considered when deciding on whether an item is functioning differentially across subpopulations. A scheme is described in a later section.

Treatment of the Polytomous Items

The polytomous items were initially treated as dichotomous, after the scores were dichotomized to "0" for scores 0 and 1 and "1" for scores 2, 3, and 4. A different dichotomization was examined with "0" for scores 0, 1, and 2 and "1" for scores 3 and 4. These dichotomizations are both arbitrary. The former gave p-values for the dichotomized polytomous items that were closer to the values of mean over the maximum score on the polytomous items than the latter. Under this treatment, a common odds ratio, log-odds and the variance of the log-odds can be computed for each polytomous item and entered into the scheme for deciding whether to flag an item for DIF or not.

There are extensions of the Mantel-Haenszel procedure for cases where the levels of a variable are more than two. In addition to their *2x2xK* analysis, Mantel and Haenszel (1959) proposed a generalized statistic for more than two response categories in a variable. A chi-square test for the case of *T* ordered response categories, where *T* can assume values larger than 2, was provided by Mantel (1963). Scores need to be assigned to each category, and a deviation of the sum of cross products from the expectation, and its variance conditioned on all marginal totals can be computed. Table 2.3 demonstrates how the data can be arranged in general and in the case of a 0-4 scored polytomous item *i* on a j^{th} partial table.

Table 2.3

Counts for the *j*th Partial Table for a Polytomous Common Item *i*

		Score on item *i*				
		0	1	…	4	Total
		Y_1	Y_2	…	Y_T	
Group	Year 1	N_{10j}	N_{11j}	…	N_{1Tj}	N_{1+j}
	Year 2	N_{20j}	N_{21j}	…	N_{2Tj}	N_{2+j}
	Total	N_{+0j}	N_{+1j}	…	N_{+Tj}	N_{++j}

According to Mantel (1963) the chi-square statistic under the null hypothesis of no association is

$$Mantel's\ \chi^2 = \frac{\left(\sum_j \sum_T N_{2Tj} Y_T - \sum_j \frac{N_{2+j}}{N_{++j}} \sum_T N_{+Tj} Y_T\right)^2}{\sum_j Var\left(\sum_T N_{2Tj} Y_T\right)} \tag{2.3}$$

The variance terms in the denominator of 2.3 are

$$Var\left(\sum_T N_{2Tj} Y_T\right) = \frac{N_{1+j} N_{2+j}}{N_{++j}^2 \left(N_{++j} - 1\right)} \left[N_{++j} \sum_T N_{+Tj} Y_T^2 - \left(\sum_T N_{+Tj} Y_T\right)^2\right] \tag{2.4}$$

Under the null hypothesis of the common odds equal to one, *Mantel's* ^{2} has a chi-square distribution with one degree of freedom. For the purposes of differential item behavior, rejecting the null hypothesis suggests that members of two subpopulations matched on a measure of proficiency differ in their mean performance on the item under investigation (Zwick, Donoghue, & Grima, 1993).

As in the case of the dichotomous items, judgments as to whether or not a polytomous item exhibits DIF take into account a measure of effect size in addition to statistical significance. Dorans and colleagues (Dorans & Kulick, 1986; Dorans & Schmitt, 1991/1993) proposed a measure of the standardized mean differences, which compares item performance of two subpopulations adjusting for differences in the distributions of the two subpopulations. Zwick, et al. (1993) reformulated the Standardized Mean Difference (SMD) as follows:

$$SMD = \sum_j \frac{N_{2+j}}{N_{2++}} \frac{\sum_T N_{2Tj} Y_T}{N_{2+j}} - \sum_j \frac{N_{2+j}}{N_{2++}} \frac{\sum_T N_{1Tj} Y_T}{N_{1+j}} \qquad (2.5)$$

The first term in the SMD is the mean performance on the item for the Year-2 group. Subtracted from that is the mean item performance for the Year-1 group weighted by the Year-2 group distribution of the matching criterion.

Zwick and Thayer (1996) provided a standard error for the SMD based on Mantel's (1963) multivariate hypergeometric model and one based on a two-multinomial model. The former performed better in their simulation study. In a comparative study by Zwick, Thayer, and Mazzeo (1997), the SMD as a descriptive index performed best among three descriptive statistics of polytomous item DIF and together with the former standard error, as good as other 5 inferential methods when the two subpopulations had the same distribution. The hypergeometric variance of the SMD (Equation 2.6) is reported in this study

$$Var(SMD) = \sum_j \left(\frac{N_{2+j}}{N_{2++}} \right)^2 \left(\frac{1}{N_{2+j}} + \frac{1}{N_{1+j}} \right)^2 Var_H \left(\sum_T N_{2Tj} Y_T \right) \qquad (2.6)$$

The variance terms are defined as in Equation 2.4.

Some Remarks on the Matching Criterion

The performance of comparable members of the two groups is contrasted to detect differential item behavior. Holland and Thayer (1988) define comparability as "identity in those measured characteristics in which examinees may differ and that are strongly related to performance on the studied item" (p.130). In the test analyzed in Study 2 all examinees that take a form in Year 1 are compared to the examinees taking the corresponding form in Year 2. Corresponding forms have a number of common items embedded in them. Matching is done based on the number correct score on a set of common items administered to both groups of examinees, summing all item scores, without rescaling the scores on the polytomous items (Zwick, et al., 1993). The number-correct score is the usual choice in studies of DIF (Welch & Miller, 1995).

Because of the use of a score on a test in which the studied item appears and which includes the score on the studied item, the Mantel-Haenszel procedure involves some circular reasoning in what it purports to do: evaluation of differential performance on an item that taps a construct after controlling for a proxy for the performance on a domain that includes the same construct. However, the choice of a total test score may be the best available matching criterion because it is a common measure that exists for all examinees; it is typically reliable as long as the test is validated for its intended purposes; and it is more reliable than individual items (Dorans & Holland, 1993). In Study 2 the score on common items was used as a matching criterion since that was the longest, common measure that both Year-1 and Year-2 subgroups had taken.

Another issue that arises in Mantel-Haenszel studies of DIF is whether the studied item should be included in the matching criterion. Holland and Thayer (1988) conjectured that when an item is analyzed for DIF it should be included in the matching criterion, but if it exhibits substantial DIF it should be excluded when examining other items. They showed that under the Rasch model when the studied item is included in the matching score the null hypothesis for the Mantel-Haenszel holds in the population; when the studied item is excluded and there is no DIF, the procedure does not behave correctly. Zwick (1990) concurred with Holland and Thayer's findings and argued that inclusion of the studied item improves the behavior of the odds ratio with more general models as well. Donoghue, Holland, and Thayer (1993) showed that inclusion of the studied item in the matching variable results in reducing the number of false positives, i.e., items that do not behave differentially being flagged as exhibiting non-zero DIF. In general, there is agreement that the studied item must be included in the matching criterion.[9]

An underlying assumption with the use of the Mantel-Haenszel procedure for the study of DIF is that the items studied are homogeneous and unidimensional (Angoff, 1993). Unidimensionality is more than an assumption for studies of DIF; it is a part of the definition of item bias (Shepard, 1982). An item that functions differentially for a group is an item that measures a construct that departs from what the matching criterion measures. If it did not, then performance of the group on that item would not

[9] Additional analyses where conducted in which the matching criterion was purified (Dorans & Holland, 1993), i.e., the items that were flagged with intermediate DIF were excluded from the calculation of the number correct score on the common items. The same items, and no other common items, were flagged in the same DIF category under the refined criterion (see Michaelides, 2005).

be expected to depart from that predicted by the group's performance on the overall criterion.

While items that are scored on a scale with multiple points, such as essay prompts or performance assessments, are considered to capture important aspects of student knowledge that are difficult to assess with more traditional testing formats, they are likely to introduce additional dimensions unrelated to the measured construct. These other dimensions may be sources of differential group performance. The dimensionality of the matching is always a concern; it is perhaps more crucial when there are polytomous items involved. Because of the more complicated nature of open-ended performance tasks, Zwick, et al. (1993) state that construct-irrelevant factors could interfere with the intended construct and lead to larger differences between groups. Nonetheless, they generalize from Holland and Thayer’s (1988) dichotomous case that polytomous items should be included in the matching variable by simply summing the scores on all dichotomous and polytomous, items. For practical reasons, this could be the only option, since tests that include polytomously scored tasks tend to have fewer items (Welch & Miller, 1995). If a matching criterion consists of very few items, the reliability of the stratification on ability will be low. Finding an appropriate matching criterion may be difficult (Dorans & Schmitt, 1991/1993), in fact impossible, if the some of the available items are not used because they are polytomous.

A Scheme for Flagging Items for Differential Behavior

The Mantel-Haenszel procedure provides information about the statistical significance of the log-odds. With large enough samples, even small departures from the null hypothesis of the common odds (or the log-odds) could result in statistical significance. A more complete judgment of whether an item behaves differentially across subpopulations would take into account both the results of hypothesis testing and the magnitude of the ratio.

The Educational Testing Service (ETS) has developed a scheme for classifying items into categories of DIF that considers the statistical significance and the magnitude of the log-odds ratio. That scheme is applied in Study 2.

ETS uses the delta metric for item difficulties by applying the transformation $\delta = 13 - 4\,\Phi^{-1}(p)$ to the p-value, p (Dorans & Holland, 1993). The log-odds ratio is transformed to the delta metric, and referred to as the Mantel-Haenszel delta difference (MH D-DIF), by

$$MH\ D - DIF = -2.35\ ln\left(\hat{\theta}_{MH}\right)$$

Note that -2.35 is an approximation of -4/1.7. The constant -4 appears in the function that transforms p-values onto the delta metric; the constant 1.7, which also appears in equations defining the 1PL, 2PL, and 3PL IRT models, serves to convert between the logit (as with log-odds) and probit (as with delta transformation) metrics.

If MH D-DIF is equal to 1.0 for an item, then that item was easier for the focal than for the reference group by one delta point. For items of moderate difficulty, a difference of one delta represents an approximate difference of 10 points on the percentage-correct scale (Zieky, 1993). Items that were easier for the reference group have a negative MH D – DIF value.

The approximate standard error for the log-odds ratio is the square root of the expression in Equation 2.2. The estimated standard error for the MH D-DIF is the log-odds standard error multiplied by 2.35 (Dorans & Holland, 1993).

$$SE(MH\ D - DIF) = 2.35\sqrt{Var[ln(\hat{\theta}_{MH})]}$$

A dichotomous item is classified into one of three categories: A, B, and C, which correspond to negligible, intermediate, and large DIF. The classification rules[10] (Dorans & Holland, 1993; Zieky, 1993) are as follows:

- A dichotomous item is classified in **Category A** if the MH D-DIF is not significantly different from zero ($p \geq 0.05$), or if its absolute value is less than 1.0.
- A dichotomous item is classified in **Category B** if the MH D-DIF is significantly different from zero, and its absolute value is at least 1.0, and its absolute value is either less than 1.5 or not significantly greater than 1.0.
- A dichotomous item is classified in **Category C** if the MH D-DIF is significantly greater than 1.0 in absolute value, and its absolute value is at least 1.5.

[10] The rules were developed empirically, based on what constitutes *large enough* difference in performance on an item by the two groups. Such a decision is judgmental.

One approach to analyzing polytomous items would be to dichotomize them by choosing a cut-point on the scoring scale and assigning a correct response to the scores above the cut, and an incorrect for the scores below. Then they can be treated as dichotomous with the same scheme.

Since statistics have been developed to deal with DIF with polytomous items, similar empirical rules exist to guide decisions on whether a polytomous item exhibits DIF or not. The rules combine statistical significance given through Mantel's chi-square statistic (Equation 2.3) and a measure for the magnitude of the difference between the performances of the two groups. The effect size is the SMD (Equation 2.5) divided by the within-group standard deviation of the studied item, pooled over the two groups. A generalization of the scheme for the dichotomous items is used in the National Assessment of Educational Progress (NAEP) to classify the polytomous items for DIF (U.S. Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics, 2001). The corresponding rules for category assignment are:

- A polytomous item is classified in **Category AA** if either Mantel's chi-square is not significantly different from zero $(p \geq 0.05)$, or if the absolute value of the effect size is less than or equal to 0.17.
- A polytomous item is classified in **Category BB** if Mantel's chi-square is significant and the absolute value of the effect size is over 0.17 and less than or equal to 0.25.
- A polytomous item is classified in **Category CC** if Mantel's chi-square is significant and the absolute value of the effect size is over 0.25 (J. Donoghue, personal communication, June 17, 2003).

Results

The Results of the Delta-plot Method

The delta-plot of the common items in the Visual and Performing Arts Grade 4 assessment appears in Figure 2.1. The delta-plot procedure flagged two dichotomous items as outliers: 4.1 and 7.1. The former was the first common item in form 4 and was easier for the Year-1 cohort (p-value=0.59, delta-value=12.07 compared to 0.49 and 13.14 respectively for the Year-2 cohort.) The latter was the first common item in form 7 and was easier for the Year-2 cohort (p-value=0.51, delta-value=12.90 compared to 0.42 and 13.82 respectively for the Year-1 cohort.)

With the delta-plot method, the decision to flag an item as an outlier is confounded with the differences in the shape of the ability distributions of the two examinee groups. The Mantel-Haenszel procedure circumvents this problem by comparing the item performance of examinees with similar proficiency scores, thus adjusting for differences in the shapes of the ability distributions of the two groups.

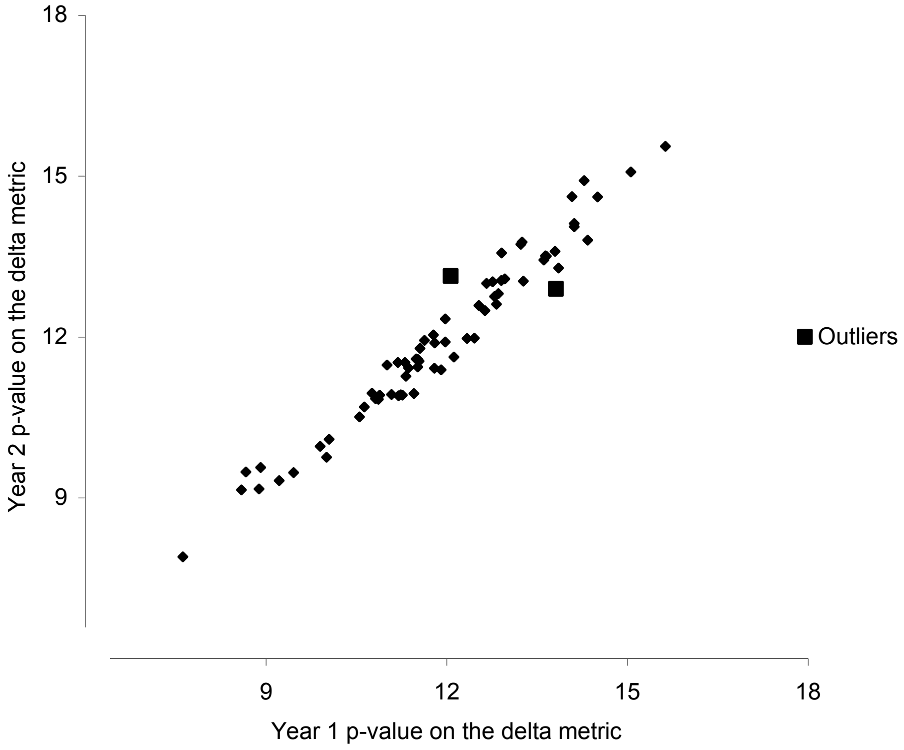


Figure 2.1. Delta-plot for the Visual and Performing Arts Grade 4 assessment.

The Results of the Mantel-Haenszel Procedure

For each common item, the relevant statistics were calculated: odds and log-odds ratio, and the standard deviation of the log-odds for the dichotomous items; for each polytomous item two analyses were carried out: one that treats it as a dichotomous item and presents the same statistics for the dichotomous case, and an alternative table with the multiple ordered response categories with the associated SMD, its standard deviation, and Mantel's chi-square statistic.

Appendix 2 lists the MH D-DIF and the associated standard errors for the sixty-nine common items of the assessment. Considering only the significance of the log-odds ratio, or equivalently of the MH D-DIF, at the 0.05 level twenty-four items would be flagged for DIF. With the application of the ETS DIF classification scheme for dichotomous items the number of the flagged items dropped considerably. Sixty-five items did not function in a significantly different way for the two cohorts and were classified in Category A. Four items, two dichotomous and two polytomous, were flagged as exhibiting intermediate or large DIF. Their statistics appear on Table 2.4. One of the polytomous items was classified in the "large DIF" category.

Table 2.4

Common Items Flagged for DIF Under the Dichotomous Treatment

Item	Item type	MH D-DIF	Standard Error (MH D-DIF)	ETS Category
5.1	Dichotomous	-1.2380	0.3131	B
7.1	Dichotomous	1.3175	0.2147	B
7.5	Polytomous	-1.4494	0.2493	B
9.6	Polytomous	-1.5405	0.2693	C

The four flagged items had an absolute value of MH D-DIF larger than one. Three of those had negative values, which means that they favored the reference, Year-1, group. The fourth item, 7.1, which was identified by the delta-plot, was easier for the focal group. Item 4.1, which was flagged by the delta-plot, was not flagged by the Mantel-Haenszel procedure; although it satisfied the significance criterion, its absolute value did not exceed the effect size threshold of one.

Table 2.5 presents results of the analysis of the polytomous data without the dichotomization. For each item, the SMD statistic, its standard deviation, the item standard deviation pooled over the two groups, the effect size which is the ratio of the SMD over the pooled standard deviation, as well as the Mantel chi-square are listed. The category in which each item is classified according to the NAEP classification scheme used by ETS appears in the last column. An SMD of 0.1 represents a difference of about one tenth of a score point in the group item means. Negative values indicate that the reference group had a higher item mean score than the focal group. The standard deviations of the SMD were very low. Item 7.5 shows a moderate difference of –0.1716 and is flagged as exhibiting intermediate DIF. Item 9.6 was flagged under the dichotomized treatment but not under the polytomous-item scheme.

Table 2.5

Statistics and Category Classifications for Polytomous Common Items

Item	SMD (SE)[1]	Item standard deviation (pooled)	Effect size	Mantel CHI-SQ	ETS Category
1.5	-0.0840 (0.0245)	0.9796	-0.0857	12.17	AA
2.6	0.0853 (0.0239)	0.8372	0.1019	12.32	AA
3.7	-0.1079 (0.0261)	0.9119	-0.1183	17.94	AA
4.6	-0.0749 (0.0248)	0.9453	-0.0792	10.30	AA
6.7	-0.0229 (0.0226)	0.7562	-0.0303	0.91	AA
7.5	-0.1716 (0.0236)	0.9489	-0.1809	53.06	BB
9.6	-0.0612 (0.0233)	0.7986	-0.0766	5.49	AA
10.5	-0.1199 (0.0260)	1.0017	-0.1197	24.36	AA
11.6	-0.0485 (0.0265)	0.9851	-0.0492	3.46	AA

[1] The standard deviation of the SMD (Equation 2.6) appears in parentheses.

Departures from unidimensionality could arise if items are measuring different skills and could result in flagging more items for DIF (Welch & Miller, 1995). To investigate the dimensionality of the data principal components, analysis of the scores on the common items within each form are presented in Appendix 3. In 19 of the 24 cases (12 forms for 2 years) only one principal component was extracted with an eigenvalue larger than one. Two principal components were extracted in 4 forms that included a polytomous item and in one form with dichotomous items only. In those cases, the loadings of items on the two principal components did not differentiate between the types of items. Finally, histograms of the number-correct score on a form showed that the distributions of the matching criterion between Year 1 and Year 2 are very similar (see Michaelides, 2003).

Discussion

Simpson's (1951) paradox is an instructive example that helps clarify the difference between the delta-plot method and the Mantel-Haenszel procedure. Dorans and Holland (1993) describe a hypothetical situation where a Group A had a higher proportion correct on an item than a Group B; however upon inspection of the performance of three strata in which the distributions of examinees were sliced (the strata could be formed by a common ability clustering in the two groups), all Group-B subgroups had higher proportion correct indices than their matched Group-A subgroups. In essence, while Group A is actually at a disadvantage as it is shown by the stratified results, the measure of overall performance implies that it is at an advantage.

When two groups do not perform equally well on an item, then the item exhibits differential impact with respect to the two groups. Impact is often stable and could replicate in other similar items, since overall ability attributes will contribute to this disparity. When the differences in ability distributions are accounted for by matching the groups on a relevant characteristic, if there are still differences between the similar subgroups, these are unexpected, given the similarity of the groups on an attribute that the item and the matching ability proxy are supposed to measure (Dorans & Holland, 1993). Matching on a relevant third variable and then comparing what is comparable has become a central concern in the study of DIF; it is crucial in making the distinction between differences in item p-values attributable to differences in item functioning versus differences in group ability (Dorans & Schmitt, 1991/1993).

Conditioning on a criterion is common to most methods of studying DIF. The delta-plot method, which was originally proposed as a technique for detecting and studying item-by-group interactions (Angoff, 1972) takes into account changes in the mean and standard deviation of the item difficulties by fitting a best-fit line. It disregards however further information about differences in the distributions of ability, and thereby confounds group differences in ability distributions with group differences in how examinees of a given ability find an item. The more the shapes of two ability distributions differ, the more the confounding is amplified when two single numbers, the p-values, are compared. It is now considered to be technically flawed for examining item bias (Angoff, 1993).

Hence, the Mantel-Haenszel procedure, as well as IRT-based methods, has taken over in studies of DIF.[11] If two administrations of a test form with common items embedded in both are considered as subgroups in a DIF-like study, then the Mantel-Haenszel procedure could provide more refined comparisons between examinees that are similar, and flag items that exhibit either homogeneous odds greater than or less than 1 across all levels of ability, or certain differential patterns of odds across ability levels.

Interpretation of Results

The content of the assessment analyzed in Study 2 could raise concerns about dimensionality. A test assessing skills on topics such as visual arts, music, and theater for children in Grade 4 includes questions that address quite different content, proficiency, or skill, especially when polytomous items are added to the common-item pool. The subjectivity involved in scoring items for creativity, cultural understanding, and aesthetics could also result in inconsistent scoring. The scorers of the Year-1 and the Year-2 administrations could be quite different in their scoring patterns, thus introducing additional dimensions in the scores.

For the particular assessment the Mantel-Haenszel procedure seemed to work well. More items were flagged as exhibiting DIF than outliers identified by the delta-plot. There was some overlap between the two methods since one of the two delta-plot outliers was flagged by the Mantel-Haenszel procedure, too. The latter flagged three additional items. Principal component analysis and histograms of ability distributions did not raise serious concerns about departures from dimensionality that could result in detecting false-positive occurrences of DIF.

The idiosyncratic features of the Visual and Performing Arts assessment might have been expected to lead to many items being flagged for DIF. However, very few items were flagged. With more common items embedded in corresponding forms the stratification of the ability variable could be more refined, matching subgroups that were even more similar in ability, increasing sensitivity of the Mantel-Haenszel procedure to actual DIF versus false-positives.

[11] IRT-based methods require IRT calibrations. In contrast, the Mantel-Haenszel procedure relies on raw scores and thus is easier to program and compute.

The delta-plot did not identify any polytomous items. The Mantel-Haenszel procedure flagged two polytomous items when polytomously scored items were dichotomized, and one item when they were treated as polytomous. Item 7.5 identified by both treatments, had a SMD of -0.1716 which suggests that there is some moderate difference in the performance (or scoring) of the two cohorts. As with all flagged items, the next step would be to inspect it through analysis of the content to detect whether it is unique compared to other items in the matching criterion. Scoring analysis of polytomous items can reveal whether the observed differential performance was due to inconsistent scoring and not due to actual examinee responses. The Year-2 scorers could re-score randomly selected responses of Year-1 examinees and compare their scoring practice with the Year-1 scores to detect any differential patterns.

Test Design and Implementation Issues

There are certain requirements on the test design that need to hold to apply the Mantel-Haenszel procedure for the study of differential behavior of common items across administrations. The Visual and Performing Arts Grade 4 assessment analyzed in this study was the only test out of more than twenty statewide assessments inspected that did not violate requirements of a Mantel-Haenszel implementation. All assessments were constructed under a matrix-sampling design; there were multiple test forms in each of two annual administrations. A form in the Year-2 administration corresponded to a form in the Year-1 administration because they shared a set of common items, the items that would equate the two annual scales. However, for purposes of linking the Year-2 to the Year-1 assessment, it is not two corresponding forms that are equated, but all alternate forms of one year to all alternate forms of the other year. What happened in most available data sets was that a common item appearing in a form X in Year 1 would be moved to a different form Y in Year 2. In essence, the common-item pools were not stable within all forms, a possibly unsound practice given the research findings on context effects reviewed earlier. If unstable common-item form assignments led to serious violations of parameter invariance assumptions due to context effects, it would be expected that larger numbers of items would be flagged, even using the delta-plot procedure. No evidence of such problems was found. In other cases, the testing program would change the number of alternate forms from one year to the next, rearranging some of the common items to newly introduced forms. Yet a few data sets that avoided moving common items around forms had as few as three or four common items in corresponding forms. Some writing assessments had only polytomous items as common. The assessment that was eventually analyzed was the only one that for all forms had a proper matching criterion, i.e., for all forms, all the common items of a Year-1 form appeared in the corresponding Year-2 form, and which consisted of a number of items that was not prohibitively low.

In typical studies of DIF these problems are not likely to emerge because the groups usually compared, gender or ethnicity groups, take the same test form, thus all test items, not just the common items, can be part of the matching criterion. In the case of equating and using Mantel-Haenszel to study the behavior of the common items, design issues are more complicated. The matching criterion can only be as long as the common-item pool, i.e., a fraction of the total test. Provided that the size of the common-item pool is large enough, the matching criterion could be refined enough to provide many strata for reliable matching. But if the common items are spread across many forms, as is the case of matrix-sampling designs, and especially if there are polytomous items in the assessment, the number of items used to form the matching variable is severely limited.

Conclusions

The choice of items to include as common in a common-item nonequivalent groups design influences the equated scores and their accuracy. In the first study it was shown that the treatment of few common items that behaved in unexpected ways across administrations, i.e., the outliers flagged by the delta-plot method, could have substantial influence on equated score summaries. In two out of four assessments analyzed, mean scores, annual gains, and proportions above a cut score differed significantly depending on whether the outlying items were included in the equating or not.

In the second study, the Mantel-Haenszel procedure, widely used in studies for identifying DIF, was proposed as an alternative to the delta-plot method and applied in the context of test equating for flagging common items that behaved differentially across cohorts of examinees. The Mantel-Haenszel procedure has the advantage of conditioning on ability when making comparisons of performance of two groups on an item. There are schemes for interpreting the effect size of differential performance, which can inform the decision as to whether to retain those items in the common-item pool, or to discard them. However, there may be some test-design limitations that preclude the application of this procedure in a test-equating framework.

Testing programs repeatedly administer different versions of the same test, and it is a matter of fairness to individual examinees to preserve comparability of the different test forms. Equating methods provide the statistical adjustments for placing scores on a common scale. Common scales are also useful for tracking trends in group performance over time, indicating how examinee cohorts perform compared to their counterparts. The significance of measuring trends accurately has recently received much attention. Under the No Child Left Behind Act of 2001, educational institutions are expected to demonstrate adequate yearly progress in reading and mathematics. There are rewards and sanctions attached to their progress. Equating is an essential part of linking achievement scores across years and maintaining a common longitudinal scale. To implement the legislation, and measure the magnitude of gains or declines from one year to the next properly, the equating must be accurate.

There are many methods to perform equating, which are usually applied within a common-item design. The effect that some misbehaving common items can have on group-level statistics–and to a much lesser degree on the variability of individual scores–is large enough to warn about the consequences of ignoring the findings from the first study, concerning variability in estimates for aggregate scores.

It is not easy to provide strict guidelines on how to deal with common items flagged for differential behavior across two forms. The content tested by a common item and its relevance to both the curriculum framework and actual instruction comes into the decision as to how to treat it, if it behaves in unexpected ways. As in the case of DIF studies where an instance of an item functioning differentially for two groups does not necessarily imply that the item is biased and should be discarded from a test (Linn, 1993), finding a common item that fails to function consistently across administrations does not imply that it is inappropriate for equating. If a common item is flagged by the Mantel-Haenszel procedure (or the delta-plot method) as behaving differentially in any two forms, it does not mean that it should be automatically discarded from the common item pool and treated as a new, non-common item in the second form. Content experts and test developers may be able to offer plausible explanations for the differential behavior. If a context effect has, for example, been discovered, then it is probably legitimate to say that it is unrelated to the construct that the test is measuring. However, as regards to equating, even in obvious cases of discrepant performance due to irrelevant circumstances, discarding a common item is not as straightforward. Common items are chosen to meet certain content and statistical specifications, and to proportionally represent the properties of the total test. Discarding a common item might violate those guidelines and introduce a different kind of bias in the equating transformation.

Even though this judgmental step is involved in the equating process, the practice can be improved in different ways. For example, the impact of including or excluding an item on the specifications and the content representation of the common-item pool can be examined. If exclusion of an item violates those specifications seriously, e.g., if it is the single item from a particular area of the tested subject, then discarding it may not be not be advisable. Another instructive piece of information is the effect that a single common item can have on the equating transformation and the equated scores. With knowledge of a common item's leverage, the decision on how to deal with it can be more informed. A third way would be the kind of information given by the schemes for flagging items for DIF implemented in Study 2. Inclusion of the effect size of the differential behavior, in addition to the statistical significance, to characterize the amount of DIF and labeling it as negligible, intermediate, or large, can be useful in deciding whether a flagged item should be discarded or not.

Limitations and Future Research

Beyond the extent of influence that misbehaving common items can have on equating results, the reasons behind the unexpected behavior are worthy of investigation. The content and the context of the common items were not examined in the aforementioned studies, although they have some bearing on the decision as to how to deal with the outliers, as has been discussed. The fact that there are very few, if any, extreme outliers in a delta-plot of common items, suggests that those outliers that do appear are probably caused by random, context and, from an educational perspective, uninteresting events, as opposed to intentional curriculum and policy changes.[12] A broad, statewide modification of the content standards for example would probably affect the behavior of many items over a long time period, while an isolated incident between two administrations that sensitized one of the examinee populations, or a change in the presentation of an item from one form to the next, would likely affect the behavior of the relevant item only. Nevertheless, outlying common items need not be the only studied items in this context. If a new policy is implemented between two administrations, then its effects can be examined on all common items, whether they exhibit large, small, or no differential behavior. Since educational policies often require longer time periods to implement and produce results, such a study could be

[12] Some causes of inconsistent item behavior across administrations can be predicted and limited. Context effects in particular are easier to control. As testing programs develop, they should incorporate sound equating habits that would avoid undesirable practices, such as changes in the format or the position of an item. Other effects probably cannot be anticipated and avoided.

longitudinal rather than just between two administrations, and would provide support for the link between educational policies and educational outcomes, reflected as changes in item (or item-cluster) performance.

The uniqueness of each testing program and the specific situations under which items are administered make it difficult to devise preset rules for dealing with misbehaving common items. As in studies of DIF, numbers by themselves cannot provide definite decision rules with regard to the complicated and sensitive issues of DIF and fairness (Zieky, 1993). Equating practice can be augmented, however, by more informative procedures. How to apply a Mantel-Haenszel procedure to flag items with a real data set has been empirically demonstrated in Study 2. Characterizing effect size of the differential behavior of items further facilitates judgments as to how to treat them. An additional useful tool would be a procedure for evaluating the leverage of each item, given its characteristics: type of item, position on a scatterplot, distance from a best-fit line, etc. Studies that simulate realistic situations would provide insight on the importance of item characteristics that affect the leverage of outliers. Vukmirovic, Hu, and Turner (2003) ran one such simulation, but defined outliers on a scatterplot of IRT *b* parameters instead of p-values on a delta-plot. Using such information together with the plausible causes of differential item behavior—content, context, or unidentifiable—the decision to keep or discard a common item can be more defensible.

Effects of outlying common items on equated scores can be examined with other equating methods, such as the IRT fixed parameter method, or non-IRT methods, in addition to the four IRT methods with equating transformations reported in Study 1. The occasional differences among methods that were observed could be investigated further. In combination with the IRT model used for test calibration there were some discrepancies between equating method results, in particular with a 3PL model for the dichotomous items. The score distributions given by the 1PL and the 3PL models were quite different in their characteristics. Comparison of the two models with respect to their scoring outcomes is another critical aspect of test analysis.

References

Allen, N. L., Ansley, T. N., & Forsyth, R. A. (1987). The effect of deleting content-related items on irt ability estimates. *Educational and Psychological Measurement, 47*(4), 1141-1152.

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

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational Measurement* (2nd edition, pp. 508-600). Washington, DC: American Council on Education. (Reprinted as W. H. Angoff, *Scales, norms, and equivalent scores.* Princeton, NJ: Educational Testing Service, 1984.)

Angoff, W. H. (1972, Sept.). *A technique for the investigation of cultural differences.* Paper presented at the annual meeting of the American Psychological Association, Honolulu. (ERIC Document Reproduction Service No. ED 069686)

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In P. W. Holland, and H. Wainer (Eds.), *Differential item functioning* (pp. 3-23). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Angoff, W. H., & Cowell, W. R. (1985). *An examination of the assumption that the equating of parallel forms is population independent* (RR-85-22). Princeton, NJ: Educational Testing Service.

Béguin, A. A. (2002, April). *Robustness of IRT test equating to violations of the representativeness of the common items in a nonequivalent groups design.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord, and M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 395-479). Reading, MA: Addison-Wesley.

Bock, R. D., Muraki, E., & Pfeiffenberger, W. (1988). Item pool maintenance in the presence of item parameter drift. *Journal of Educational Measurement, 25*(4), 275-285.

Brennan, R. L., & Kolen, M. J. (1987). Some practical issues in equating. *Applied Psychological Measurement, 11*(3), 279-290.

Budescu, D. (1985). Efficiency of linear equating as a function of the length of the anchor test. *Journal of Educational Measurement,22*(1), 13-20.

Cassels, J. R. T., & Johnstone, A. H. (1984). The effect of language on student performance on multiple choice tests in chemistry. *Journal of Chemical Education, 61*, 613-615.

Cizek, G. J. (1994). The effect of altering the position of options in a multiple-choice examination. *Educational and Psychological Measurement, 54*(1), 8-20.

Cook, L. L., Eignor, D. R., & Taft, H. L. (1988). A comparative study of the effects of recency of instruction on the stability of irt and conventional item parameter estimates. *Journal of Educational Measurement, 25*(1), 31-45.

Cook, L. L., & Petersen, N. S. (1987). Problems related to the use of conventional and item response theory equating methods in less than optimal circumstances. *Applied Psychological Measurement, 11*(3), 225-244.

Donoghue, J. R., Holland, P. W., & Thayer, D. T. (1993). A Monte Carlo study of factors that affect the mantel-haenszel and standardization measures of differential item functioning. In P. W. Holland, and H. Wainer (Eds.), *Differential item functioning* (pp. 137-166). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland, and H. Wainer (Eds.), *Differential item functioning* (pp. 35-66). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Dorans, N. J., & Holland, P. W. (2000). Population invariance and the equitability of tests: basic theory and the linear case. *Journal of Educational Measurement, 37*(4), 281-306.

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement, 23,* 355-368.

Dorans, N. J., & Schmitt, A. P. (1991). *Constructed-response and differential item functioning: A pragmatic approach* (ETS Research Report No. 91-47). Princeton, NJ: Educational Testing Service. [Also appears in *Construction vs. choice in cognitive measurement* (pp. 135-166), R. E. Bennett & W. C. Ward (Eds.), 1993, Hillsdale, NJ: Erlbaum.]

Gilmer, J. S. (1989). The effects of test disclosure on equated scores and pass rates. *Applied Psychological Measurement, 13*, 245-255.

Goldstein, H. (1980). Dimensionality, bias, independence and measurement scale problems in latent trait test score models. *British Journal of Mathematical and Statistical Psychology, 33,* 234-246.

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research, 22,* 144-149.

Hambleton, R. K., Swaminathan, H., & Rogers H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage Publications, Inc.

Hanson, B. A., & Feinstein, Z. S. (1997). *Application of a polynomial loglinear model to assessing differential item functioning for common items in the common-item equating design* (ACT Research Report Series No. 97-1). Iowa City, IA: ACT Inc.

Hanson, B. A., & Zeng, L. (1995). ST. A computer program for IRT equating (Version 1.0) [Computer software and manual]. Retrieved from http://www.uiowa.edu/~itp/pages/SWEQUATING.SHTML

Hattie, J., Jaeger, R. M., & Bond, L. (1999). Persistent methodological questions in educational testing. *Review of Research in Education, 23*, 393-446.

Holland, P. W. (1985). On the study of differential item performance without IRT. *Proceedings of the Military Testing Association*, October.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer, and H. I. Brown (Eds.), *Test validity* (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum Associates.

Huang, C.-Y., & Shyu, C.-Y. (2003, April). *The impact of item parameter drift on equating*. Paper presented at the annual meeting of the National Council on Measurement in Education. Chicago, Il.

Hulin, C. L., Lissak, R. I., & Drasgow, F. (1982). Recovery of two- and three-parameter logistic item characteristic curves: A Monte Carlo study. *Applied Psychological Measurement, 6*(3), 249-260.

Klein, L. W., & Jarjoura, D. (1985). The importance of content representation for common-item equating with nonrandom groups, *Journal of Educational Measurement, 22*(3), 197-206.

Kingston, N. M., & Dorans, N. J. (1984). Item location effects and their implications for IRT equating and adaptive testing. *Applied Psychological Measurement*, *8*, 147-154.

Kingston, N., Leary, L., & Wightman, L. (1985). *An exploratory study of the applicability of item response theory methods to the Graduate Management Admissions Test* (RR-85-34). Princeton, NJ: Educational Testing Service.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking methods and practices* (2nd ed.). New York: Springer.

Linn, R. L. (1990). Has item response theory increased the validity of achievement test scores? *Applied Measurement in Education, 3*(2), 115-141.

Linn, R. L. (1993). The use of differential item functioning statistics: A discussion of current practice and future implications. In P. W. Holland, and H. Wainer (Eds.), *Differential item functioning* (pp. 349-364). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

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

Loyd, B. H., & Hoover, H. D. (1980). Vertical equating using the rasch model. *Journal of Educational Measurement, 17*(3), 179-193.

Mantel, N. (1963). Chi-square tests with one degree of freedom; Extensions of the Mantel-Haenszel procedure. *Journal of the American Statistical Association, 58*, 690-700.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22*, 719-748.

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement, 14*(2), 139-160.

Masters, G.N. (1988). Item discrimination: When more is worse. *Journal of Educational Measurement, 25*(1), 15-29.

Mehrens, W. A., & Phillips, S. E. (1986). Detecting impacts of curricular differences in achievement test data. *Journal of Educational Measurement, 23*(3), 185-196.

Mehrens, W. A., & Phillips, S. E. (1987). Sensitivity of item difficulties to curricular validity. *Journal of Educational Measurement, 24*(4), 357-370.

Michaelides, M. P. (2003). *Effects of common-item selection on the accuracy of item response theory test equating with nonequivalent groups*. Unpublished doctoral dissertation. Stanford University, Stanford, CA.

Michaelides, M. P. (2005). *An application of a Mantel-Haenszel procedure to flag misbehaving common items in test equating*. Poster presented at the Annual Meeting American Educational Research Association, Montreal, Canada.

Miller, A. D., & Linn, R. L. (1988). Invariance of item characteristic functions with variations in instructional coverage. *Journal of Educational Measurement, 25*(3), 205-219.

Mitzel, H. C., Weber, M. M., & Sykes, R. C. (1999). *Test item disclosure: How much difference does it really make?* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Montreal, Canada.

Muraki, E., & Bock, R. D. (1997). PARSCALE (Version 3.0) [Computer software]. Chicago, Il.: Scientific Software International, Inc.

National Research Council. (1999). *Embedding questions: The pursuit of a common measure in uncommon tests.* Committee on Embedding Common Test Items in State and District Assessments. D. M. Koretz, M. W. Bertenthal, and B. F. Green, eds. Board on Testing and Assessment, Commission on Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.

Phillips, A., & Holland, P. W. (1987). Estimators of the variance of the Mantel-Haenszel log-odds-ratio estimate. *Biometrics, 43,* 425-431.

Phillips, S. E., & Mehrens, W. A. (1987). Curricular differences and unidimensionality of achievement test data: An exploratory analysis. *Journal of Educational Measurement, 24*(1), 1-16.

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores* (Psychometric Monograph No. 17). Iowa City, IA: Psychometric Society.

Shepard, L. A. (1982). Definitions of bias. In R. A. Berk (Ed.), *Handbook of methods for detecting test bias* (pp. 9-30). Baltimore: John Hopkins University Press.

Simpson, E. H. (1951). Interpretation of interaction contingency tables. *Journal of the Royal Statistical Society, (Series B), 13*, 238-241.

Skaggs, G., & Lissitz, R. W. (1986). IRT test equating: Relevant issues and a review of recent research. *Review of Educational Research, 56*(4), 495-529.

Stahl, J., Bergstrom, B., & Shneyderman, O. (2002, April). *Impact of item drift on person measurement*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Journal of Educational Measurement, 7*, 201-210.

Sykes, R. C., & Fitzpatrick, A. R. (1992). The stability of IRT *b* values. *Journal of Educational Measurement, 29*(3), 201-211.

Traub, R. E. (1983). A priori considerations in choosing an item response model. In R. K. Hambleton (Ed.), *Applications of item response theory* (pp. 57-70). Vancouver, BC: Educational Research Institute of British Columbia.

U.S. Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics. *The NAEP 1998 technical report*, NCES 2001-509, by Allen, N. L., Donoghue, J. R., & Schoeps, T. L. (2001). Washington, DC: National Center for Education Statistics.

Vukmirovic, Z., Hu, H., & Turner, J. C. (2003, April). The effects of outliers on IRT equating with fixed common item parameters. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, Il.

Wainer, H. (1999). Comparing the incomparable: An essay on the importance of big assumptions and scant evidence. *Educational Measurement: Issues and Practice, 18*(4), 10-16.

Way, W. D., Forsyth, R. A., & Ansley, T. N. (1989). IRT ability estimates from customized achievement tests without representative content sampling. *Applied Measurement in Education, 2*(1), 15-35.

Welch, C. J., & Miller, T. R. (1995). assessing differential item functioning in direct writing assessments: Problems and an example. *Journal of Educational Measurement, 32*(2), 163-178.

Wells, C. S., Subkoviak, M. J., & Serlin, R. C. (2002). The effect of item parameter drift on examinee ability estimates. *Applied Psychological Measurement, 26*(1), 77-87.

Yen, W. M. (1980). The extent, causes and importance of context effects on item parameters for two latent trait models. *Journal of Educational Measurement, 17*(4), 297-311.

Yen, W. M., Green, D. R., & Burket, G. R. (1987). Valid information from customized achievement tests. *Educational Measurement: Issues and Practice, 6*(1), 7-13.

Zieky, M. (1993). Practical questions in the use of dif statistics in test development. In P. W. Holland, and H. Wainer (Eds.), *Differential item functioning* (pp. 337-347). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Zwick, R. (1990). When do item response function and Mantel-Haenszel definitions of differential item functioning coincide? *Journal of Educational Statistics, 15*(3), 185-197.

Zwick, R., Donoghue, J. R., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement, 30*(3), 233-251.

Zwick, R., & Thayer, D. T. (1996). Evaluating the magnitude of differential item functioning in polytomous items. *Journal of Educational and Behavioral Statistics, 21*(3), 187-201.

Zwick, R., Thayer, D. T., & Mazzeo, J. (1997). Descriptive and inferential procedures for assessing differential item functioning in polytomous items. *Applied Measurement in Education, 10*(4), 321-344.

APPENDIX 1: ST Output – Equating Transformations

Appendix 1 presents item parameter statistics and equating transformation constants:

(a) under four equating methods: Stocking & Lord, Haebara, mean/mean, and mean/sigma as given by the ST software (Hanson & Zeng, 1995),

(b) for each of the four assessments analyzed in study 1: Mathematics 8, Science 11, Social Studies 6, and Science 6,

(c) under a 3PL IRT and a 1PL IRT calibration, and

(d) using all versus using only the non-outlying common items.

MATHEMATICS 8, 3PL, ALL COMMON ITEMS

```
ST 1.0.2
Input file: c:\data\equate\param\8m_all_st_in.txt

Number of common items: 65

   Item Parameter Means
      a    b    c
New Form   0.8713  0.6970  0.1033
Old Form   0.9830  0.4868  0.1033

 Item Parameter Standard Deviations
      a    b    c
New Form   0.2718  1.3984  0.0973
Old Form   0.2700  1.1832  0.0973

Transformation Functions

    Stocking-Lord  Haebara  Mean/Mean Mean/Sigma
Intercept   -0.119995  -0.134256  -0.131011  -0.102949
Slope     0.883241  0.896178  0.886359  0.846100
```

MATHEMATICS 8, 3PL, NON-OUTLYING COMMON ITEMS

```
ST 1.0.2
Input file: c:\data\equate\param\8m_NONOUTLIERS_st_in.txt

Number of common items: 62

   Item Parameter Means
      a    b    c
New Form   0.8687  0.7227  0.1001
Old Form   0.9879  0.5036  0.1001

 Item Parameter Standard Deviations
      a    b    c
New Form   0.2633  1.4259  0.0980
Old Form   0.2717  1.2084  0.0980

Transformation Functions

    Stocking-Lord  Haebara  Mean/Mean Mean/Sigma
Intercept   -0.127278  -0.142774  -0.131884  -0.108832
Slope     0.881048  0.897917  0.879361  0.847461
```

SCIENCE 11, 3PL, ALL COMMON ITEMS

```
ST 1.0.2
Input file: C:\DATA\EQUATE\PARAM\11C_ALL_ST_IN.TXT

Number of common items: 63

   Item Parameter Means
      a    b    c
New Form   0.8020  0.5610  0.1121
Old Form   0.8587  0.4271  0.1121

 Item Parameter Standard Deviations
      a    b    c
New Form   0.1987  1.0474  0.1029
Old Form   0.2295  0.9324  0.1029

Transformation Functions

     Stocking-Lord  Haebara  Mean/Mean Mean/Sigma
Intercept   -0.085172  -0.072817  -0.096918  -0.072329
Slope     0.956795  0.953115  0.933978  0.890149
```

SCIENCE 11, 3PL, NON-OUTLYING COMMON ITEMS

```
ST 1.0.2
Input file: C:\DATA\EQUATE\PARAM\11C_NONOUTLIERS_ST_IN.TXT

Number of common items: 54

   Item Parameter Means
      a    b    c
New Form   0.8002  0.5970  0.1263
Old Form   0.8300  0.4024  0.1263

 Item Parameter Standard Deviations
      a    b    c
New Form   0.2133  0.9841  0.0999
Old Form   0.2343  0.8579  0.0999

Transformation Functions

     Stocking-Lord  Haebara  Mean/Mean Mean/Sigma
Intercept   -0.146993  -0.138975  -0.173123  -0.117985
Slope     0.971272  0.955871  0.964080  0.871721
```

SOCIAL STUDIES 6, 3PL, ALL COMMON ITEMS

```
ST 1.0.2
Input file: C:\DATA\EQUATE\PARAM\6S_ALL_ST_IN.TXT

Number of common items: 89

   Item Parameter Means
      a    b    c
New Form   0.7503  0.1937  0.0732
Old Form   0.7620  0.2982  0.0732

 Item Parameter Standard Deviations
      a    b    c
New Form   0.1526  1.3775  0.0793
Old Form   0.1503  1.3350  0.0793

Transformation Functions

    Stocking-Lord  Haebara  Mean/Mean Mean/Sigma
Intercept   0.110536  0.111957  0.107489  0.110490
Slope     0.979559  0.974644  0.984664  0.969167
```

SOCIAL STUDIES 6, 3PL, NON-OUTLYING COMMON ITEMS

```
ST 1.0.2
Input file: C:\DATA\EQUATE\PARAM\6S_NONOUTLIERS_ST_IN.TXT

Number of common items: 87

   Item Parameter Means
      a    b    c
New Form   0.7514  0.2013  0.0725
Old Form   0.7629  0.3056  0.0725

 Item Parameter Standard Deviations
      a    b    c
New Form   0.1527  1.3923  0.0800
Old Form   0.1512  1.3492  0.0800

Transformation Functions

    Stocking-Lord  Haebara  Mean/Mean Mean/Sigma
Intercept   0.110608  0.112376  0.107311  0.110518
Slope     0.979903  0.974575  0.984975  0.969038
```

SCIENCE 6, 3PL, ALL COMMON ITEMS

```
ST 1.0.2
Input file: C:\DATA\EQUATE\PARAM\6C_ALL_ST_IN.TXT

Number of common items: 80

   Item Parameter Means
     a    b    c
New Form   0.6625  0.3067  0.0815
Old Form   0.7259  0.2987  0.0815

 Item Parameter Standard Deviations
     a    b    c
New Form   0.1542  1.5461  0.0895
Old Form   0.1650  1.3963  0.0895

Transformation Functions

    Stocking-Lord  Haebara  Mean/Mean Mean/Sigma
Intercept   0.048868  0.085721  0.018857  0.021818
Slope     0.930364  0.951653  0.912738  0.903081
```

SCIENCE 6, 3PL, NON-OUTLYING COMMON ITEMS

```
ST 1.0.2
Input file: C:\DATA\EQUATE\PARAM\6C_NONOUTLIERS_ST_IN.TXT

Number of common items: 79

   Item Parameter Means
     a    b    c
New Form   0.6637  0.3109  0.0807
Old Form   0.7274  0.2979  0.0807

 Item Parameter Standard Deviations
     a    b    c
New Form   0.1549  1.5554  0.0898
Old Form   0.1655  1.4051  0.0898

Transformation Functions

    Stocking-Lord  Haebara  Mean/Mean Mean/Sigma
Intercept   0.044031  0.081835  0.014229  0.017038
Slope     0.929829  0.951510  0.912375  0.903341
```

MATHEMATICS 8, 1PL, ALL COMMON ITEMS

```
ST 1.0.2
Input file: C:\DATA\EQUATE\PARAM\8M_1PL_ALL_ST_IN.TXT

Number of common items: 65

   Item Parameter Means
      a    b    c
New Form   1.0000  0.3117  0.0000
Old Form   1.0000  0.3356  0.0000

 Item Parameter Standard Deviations
      a    b    c
New Form   0.0000  0.8343  0.0000
Old Form   0.0000  0.8410  0.0000

Transformation Functions

    Stocking-Lord  Haebara  Mean/Mean Mean/Sigma
Intercept   0.028459  0.028299  0.023941  0.021442
Slope     1.017464  1.026623  1.000000  1.008017
```

MATHEMATICS 8, 1PL, NON-OUTLYING COMMON ITEMS

```
ST 1.0.2
Input file: C:\DATA\EQUATE\PARAM\8M_1PL_NONOUTLIERS_ST_IN.TXT

Number of common items: 62

   Item Parameter Means
      a    b    c
New Form   1.0000  0.3323  0.0000
Old Form   1.0000  0.3558  0.0000

 Item Parameter Standard Deviations
      a    b    c
New Form   0.0000  0.8488  0.0000
Old Form   0.0000  0.8557  0.0000

Transformation Functions

    Stocking-Lord  Haebara  Mean/Mean Mean/Sigma
Intercept   0.028158  0.027924  0.023530  0.020835
Slope     1.018189  1.027399  1.000000  1.008111
```

SCIENCE 11, 1PL, ALL COMMON ITEMS

```
ST 1.0.2
Input file: C:\DATA\EQUATE\PARAM\11C_1PL_ALL_ST_IN.TXT

Number of common items: 63

   Item Parameter Means
      a    b    c
New Form   1.0000  0.2334  0.0000
Old Form   1.0000  0.2858  0.0000

 Item Parameter Standard Deviations
      a    b    c
New Form   0.0000  0.8303  0.0000
Old Form   0.0000  0.8325  0.0000

Transformation Functions

    Stocking-Lord  Haebara  Mean/Mean Mean/Sigma
Intercept   0.047937  0.043177  0.052410  0.051799
Slope    1.005713  1.020442  1.000000  1.002619
```

SCIENCE 11, 1PL, NON-OUTLYING COMMON ITEMS

```
ST 1.0.2
Input file: C:\DATA\EQUATE\PARAM\11C_1PL_NONOUTLIERS_ST_IN.TXT

Number of common items: 54

   Item Parameter Means
      a    b    c
New Form   1.0000  0.2409  0.0000
Old Form   1.0000  0.2395  0.0000

 Item Parameter Standard Deviations
      a    b    c
New Form   0.0000  0.7974  0.0000
Old Form   0.0000  0.7548  0.0000

Transformation Functions

    Stocking-Lord  Haebara  Mean/Mean Mean/Sigma
Intercept   0.011760  0.014734  -0.001398  0.011484
Slope    0.993381  0.985601  1.000000  0.946528
```

SOCIAL STUDIES 6, 1PL, ALL COMMON ITEMS

```
ST 1.0.2
Input file: C:\DATA\EQUATE\PARAM\6S_1PL_ALL_ST_IN.TXT

Number of common items: 89

   Item Parameter Means
      a    b    c
New Form   1.0000  0.2040  0.0000
Old Form   1.0000  0.1880  0.0000

 Item Parameter Standard Deviations
      a    b    c
New Form   0.0000  0.9950  0.0000
Old Form   0.0000  0.9965  0.0000

Transformation Functions

    Stocking-Lord  Haebara  Mean/Mean Mean/Sigma
Intercept   -0.014749  -0.011598  -0.016018  -0.016331
Slope     0.999557  0.991314  1.000000  1.001538
```

SOCIAL STUDIES 6, 1PL, NON-OUTLYING COMMON ITEMS

```
ST 1.0.2
Input file: C:\DATA\EQUATE\PARAM\6S_1PL_NONOUTLIERS_ST_IN.TXT

Number of common items: 87

   Item Parameter Means
      a    b    c
New Form   1.0000  0.2096  0.0000
Old Form   1.0000  0.1933  0.0000

 Item Parameter Standard Deviations
      a    b    c
New Form   0.0000  1.0057  0.0000
Old Form   0.0000  1.0073  0.0000

Transformation Functions

    Stocking-Lord  Haebara  Mean/Mean Mean/Sigma
Intercept   -0.015205  -0.012009  -0.016393  -0.016731
Slope     0.999516  0.991176  1.000000  1.001613
```

SCIENCE 6, 1PL, ALL COMMON ITEMS

```
ST 1.0.2
Input file: C:\DATA\EQUATE\PARAM\6C_1PL_ALL_ST_IN.TXT

Number of common items: 80

   Item Parameter Means
       a    b    c
New Form   1.0000  0.1525  0.0000
Old Form   1.0000  0.1644  0.0000

 Item Parameter Standard Deviations
       a    b    c
New Form   0.0000  0.9497  0.0000
Old Form   0.0000  0.9412  0.0000

Transformation Functions

    Stocking-Lord  Haebara  Mean/Mean Mean/Sigma
Intercept   0.008022  0.004502  0.011876  0.013243
Slope    1.012413  1.030590  1.000000  0.991040
```

SCIENCE 6, 1PL, NON-OUTLYING COMMON ITEMS

```
ST 1.0.2
Input file: C:\DATA\EQUATE\PARAM\6C_1PL_NONOUTLIERS_ST_IN.TXT

Number of common items: 79

   Item Parameter Means
       a    b    c
New Form   1.0000  0.1556  0.0000
Old Form   1.0000  0.1664  0.0000

 Item Parameter Standard Deviations
       a    b    c
New Form   0.0000  0.9554  0.0000
Old Form   0.0000  0.9470  0.0000

Transformation Functions

    Stocking-Lord  Haebara  Mean/Mean Mean/Sigma
Intercept   0.006575  0.002854  0.010801  0.012160
Slope    1.012654  1.031380  1.000000  0.991272
```

APPENDIX 2: Mantel-Haenszel Procedure and ETS Scheme for Flagging Dichotomous Items for DIF

FORM	ITEM	POLYTOMOUS	MH D-DIF	SE (MH D-DIF)	ETS CATEGORY
1	1		0.2700	0.2088	A
	2		0.2385	0.2796	A
	3		0.1447	0.2162	A
	4		0.4703	0.2051	A
	5	POLYTOMOUS[1]	-0.5566	0.2466	A
2	1		-0.3680	0.2132	A
	2		-0.5767	0.2213	A
	3		-0.0700	0.2110	A
	4		-0.3526	0.2196	A
	5		0.2477	0.2074	A
	6	POLYTOMOUS	0.5820	0.2493	A
3	1		-0.2969	0.3500	A
	2		0.3344	0.2202	A
	3		0.4139	0.2214	A
	4		-0.5456	0.2956	A
	5		0.3910	0.2101	A
	6		0.6510	0.2128	A
	7	POLYTOMOUS	-0.8932	0.2325	A
4	1		-0.6997	0.2153	A
	2		0.7030	0.2453	A
	3		0.7346	0.2356	A
	4		-0.3773	0.2792	A
	5		0.9568	0.2447	A
	6	POLYTOMOUS	-0.9968	0.2467	A
5	1		-1.2380	0.3131	B
	2		0.2152	0.2254	A
	3		0.1892	0.2513	A
	4		0.2934	0.2475	A
	5		0.0543	0.2444	A
6	1		-0.1766	0.2505	A
	2		-0.5405	0.2741	A
	3		0.5747	0.2227	A
	4		-0.1526	0.2369	A
	5		-0.0277	0.2257	A

	6		0.4206	0.2350	A
	7	POLYTOMOUS	-0.0803	0.2723	A
7	1		1.3175	0.2147	B
	2		0.5587	0.2337	A
	3		0.0788	0.2278	A
	4		0.5504	0.2709	A
	5	POLYTOMOUS	-1.4494	0.2493	B
8	1		-0.5309	0.2227	A
	2		0.4051	0.2170	A
	3		-0.3416	0.2794	A
	4		0.4061	0.2231	A
	5		0.1936	0.2218	A
	6		-0.3080	0.2343	A
9	1		-0.0846	0.2149	A
	2		0.1191	0.2342	A
	3		0.3330	0.2125	A
	4		-0.0793	0.2123	A
	5		0.3961	0.2086	A
	6	POLYTOMOUS	-1.5405	0.2693	C
10	1		0.3003	0.2130	A
	2		0.4034	0.2164	A
	3		0.1308	0.2092	A
	4		0.7631	0.2196	A
	5	POLYTOMOUS	-0.8708	0.2747	A
11	1		0.4784	0.2052	A
	2		-0.2745	0.2430	A
	3		0.1310	0.2024	A
	4		-0.3016	0.2546	A
	5		0.3624	0.2073	A
	6	POLYTOMOUS	-0.5489	0.2386	A
12	1		-0.2703	0.2370	A
	2		-0.0749	0.2200	A
	3		-0.3764	0.2374	A
	4		0.3104	0.2409	A
	5		0.4108	0.2310	A

[1] Polytomous items scored 0-4 are dichotomized to incorrect if score is 0 or 1 and correct if the score is 2, 3, or 4

APPENDIX 3: SPSS Output for Principal Components of Responses to Common Items by Form and Year

Appendix 3 includes the results of principal components analysis on the common-item responses for each form per year. The eigenvalues, the percentage of variance explained by each principal component, and the loadings of each item on the extracted principal components are presented.

Form 1 Year 1

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	1.537	30.748	30.748	1.537	30.748	30.748
2	.963	19.253	50.000			
3	.892	17.832	67.833			
4	.830	16.594	84.426			
5	.779	15.574	100.000			

Extraction Method: Principal Component Analysis.

Component Matrix(a)

	Component
	1
F1I1Y1	.516
F1I2Y1	.600
F1I3Y1	.603
F1I4Y1	.376
F1I5PY1	.637

Extraction Method: Principal Component Analysis.
a 1 components extracted.

Form 1 Year 2

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	1.455	29.097	29.097	1.455	29.097	29.097
2	.949	18.986	48.082			
3	.902	18.034	66.117			
4	.859	17.174	83.291			
5	.835	16.709	100.000			

Extraction Method: Principal Component Analysis.

Component Matrix(a)

	Component
	1
F1I1Y2	.473
F1I2Y2	.578
F1I3Y2	.542
F1I4Y2	.484
FII4PY2	.607

Extraction Method: Principal Component Analysis.
a 1 components extracted.

Form 2 Year 1

Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	1.691	28.175	28.175	1.691	28.175	28.175
2	.991	16.520	44.696			
3	.891	14.843	59.539			
4	.830	13.834	73.373			
5	.805	13.409	86.782			
6	.793	13.218	100.000			

Extraction Method: Principal Component Analysis.

Component Matrix(a)

	Component
	1
F2I1Y1	.513
F2I2Y1	.500
F2I3Y1	.541
F2I4Y1	.562
F2I5Y1	.455
F2I6PY1	.602

Extraction Method: Principal Component Analysis.
a 1 components extracted.

Form 2 Year 2

Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	1.542	25.699	25.699	1.542	25.699	25.699
2	1.009	16.818	42.517	1.009	16.818	42.517
3	.921	15.352	57.870			
4	.903	15.057	72.927			
5	.867	14.443	87.370			
6	.758	12.630	100.000			

Extraction Method: Principal Component Analysis.

Component Matrix(a)

	Component	
	1	2
F2I1Y2	.571	-.220
F2I2Y2	.480	-.451
F2I3Y2	.481	.470
F2I4Y2	.526	-.228
F2I5Y2	.386	.694
F2I6PY2	.573	-.056

Extraction Method: Principal Component Analysis.
a 2 components extracted.

Form 3 Year 1

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	1.838	26.254	26.254	1.838	26.254	26.254
2	1.018	14.537	40.791	1.018	14.537	40.791
3	.948	13.536	54.327			
4	.886	12.659	66.986			
5	.804	11.481	78.467			
6	.777	11.094	89.560			
7	.731	10.440	100.000			

Component Matrix(a)

	Component	
	1	2
F3I1Y1	.466	-.603
F3I2Y1	.504	.119
F3I3Y1	.602	.314
F3I4Y1	.564	-.471
F3I5Y1	.390	.511
F3I6Y1	.445	.242
F3I7PY1	.580	-.015

a 2 components extracted.

Form 3 Year 2

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	1.650	23.569	23.569	1.650	23.569	23.569
2	1.045	14.930	38.498	1.045	14.930	38.498
3	.958	13.684	52.182			
4	.893	12.751	64.933			
5	.847	12.104	77.036			
6	.835	11.934	88.970			
7	.772	11.030	100.000			

Component Matrix(a)

	Component	
	1	2
F3I1Y2	.410	-.573
F3I2Y2	.566	-.085
F3I3Y2	.595	.151
F3I4Y2	.505	-.460
F3I5Y2	.345	.443
F3I6Y2	.463	.524
F3I7PY2	.468	.064

a 2 components extracted.

Form 4 Year 1

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	1.789	29.809	29.809	1.789	29.809	29.809
2	.974	16.230	46.039			
3	.901	15.025	61.064			
4	.823	13.724	74.788			
5	.785	13.081	87.869			
6	.728	12.131	100.000			

Extraction Method: Principal Component Analysis.

Component Matrix(a)

	Component
	1
F4I1Y1	.530
F4I2Y1	.601
F4I3Y1	.308
F4I4Y1	.527
F4I5Y1	.630
F4I6PY1	.613

Extraction Method: Principal Component Analysis.
a 1 components extracted.

Form 4 Year 2

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	1.714	28.565	28.565	1.714	28.565	28.565
2	.988	16.463	45.028			
3	.914	15.227	60.255			
4	.848	14.139	74.393			
5	.811	13.516	87.909			
6	.725	12.091	100.000			

Extraction Method: Principal Component Analysis.

Component Matrix(a)

	Component
	1
F4I1Y2	.537
F4I2Y2	.577
F4I3Y2	.297
F4I4Y2	.520
F4I5Y2	.643
F4I6PY2	.566

Extraction Method: Principal Component Analysis.
a 1 components extracted.

Form 5 Year 1

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	1.780	35.603	35.603	1.780	35.603	35.603
2	.939	18.772	54.375			
3	.809	16.176	70.550			
4	.750	15.001	85.551			
5	.722	14.449	100.000			

Extraction Method: Principal Component Analysis.

Component Matrix(a)

	Component
	1
F5I1Y1	.639
F5I2Y1	.373
F5I3Y1	.646
F5I4Y1	.675
F5I5Y1	.600

Extraction Method: Principal Component Analysis.
a 1 components extracted.

Form 5 Year 2

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	1.732	34.640	34.640	1.732	34.640	34.640
2	.888	17.767	52.407			
3	.852	17.031	69.437			
4	.790	15.800	85.237			
5	.738	14.763	100.000			

Extraction Method: Principal Component Analysis.

Component Matrix(a)

	Component
	1
F5I1Y2	.587
F5I2Y2	.506
F5I3Y2	.590
F5I4Y2	.615
F5I5Y2	.637

Extraction Method: Principal Component Analysis.
a 1 components extracted.

Form 6 Year 1

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	1.928	27.543	27.543	1.928	27.543	27.543
2	.959	13.703	41.246			
3	.921	13.153	54.399			
4	.862	12.319	66.719			
5	.821	11.729	78.448			
6	.768	10.979	89.427			
7	.740	10.573	100.000			

Component Matrix(a)

	Component
	1
F6I1Y1	.513
F6I2Y1	.410
F6I3Y1	.478
F6I4Y1	.551
F6I5Y1	.556
F6I6Y1	.634
F6I7PY1	.503

a 1 components extracted.

Form 6 Year 2

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	1.837	26.248	26.248	1.837	26.248	26.248
2	.999	14.277	40.526			
3	.925	13.216	53.742			
4	.856	12.232	65.974			
5	.838	11.971	77.945			
6	.805	11.505	89.450			
7	.738	10.550	100.000			

Component Matrix(a)

	Component
	1
F6I1Y2	.508
F6I2Y2	.309
F6I3Y2	.536
F6I4Y2	.561
F6I5Y2	.574
F6I6Y2	.613
F6I7PY2	.420

a 1 components extracted.

Form 7 Year 1

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	1.777	35.539	35.539	1.777	35.539	35.539
2	.896	17.924	53.462			
3	.837	16.739	70.201			
4	.787	15.739	85.940			
5	.703	14.060	100.000			

Extraction Method: Principal Component Analysis.

Component Matrix(a)

	Component
	1
F7I1Y1	.481
F7I2Y1	.603
F7I3Y1	.641
F7I4Y1	.647
F7I5PY1	.594

Extraction Method: Principal Component Analysis.
a 1 components extracted.

Form 7 Year 2

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	1.737	34.744	34.744	1.737	34.744	34.744
2	.885	17.700	52.444			
3	.858	17.153	69.597			
4	.787	15.739	85.336			
5	.733	14.664	100.000			

Extraction Method: Principal Component Analysis.

Component Matrix(a)

	Component
	1
F7I1Y2	.503
F7I2Y2	.580
F7I3Y2	.610
F7I4Y2	.640
F7I5PY2	.605

Extraction Method: Principal Component Analysis.
a 1 components extracted.

Form 8 Year 1

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	1.586	26.429	26.429	1.586	26.429	26.429
2	1.001	16.679	43.108	1.001	16.679	43.108
3	.907	15.120	58.228			
4	.897	14.951	73.179			
5	.833	13.887	87.065			
6	.776	12.935	100.000			

Extraction Method: Principal Component Analysis.

Component Matrix(a)

	Component	
	1	2
F8I1Y1	.484	-.157
F8I2Y1	.383	.685
F8I3Y1	.643	.083
F8I4Y1	.521	.249
F8I5Y1	.458	-.649
F8I6Y1	.557	-.129

Extraction Method: Principal Component Analysis.
a 2 components extracted.

Form 8 Year 2

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	1.541	25.691	25.691	1.541	25.691	25.691
2	.995	16.591	42.282			
3	.953	15.880	58.162			
4	.909	15.157	73.319			
5	.835	13.916	87.234			
6	.766	12.766	100.000			

Extraction Method: Principal Component Analysis.

Component Matrix(a)

	Component
	1
F8I1Y2	.454
F8I2Y2	.459
F8I3Y2	.627
F8I4Y2	.594
F8I5Y2	.309
F8I6Y2	.531

Extraction Method: Principal Component Analysis.
a 1 components extracted.

Form 9 Year 1

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	1.599	26.650	26.650	1.599	26.650	26.650
2	.977	16.276	42.927			
3	.949	15.815	58.741			
4	.883	14.710	73.452			
5	.810	13.502	86.953			
6	.783	13.047	100.000			

Extraction Method: Principal Component Analysis.

Component Matrix(a)

	Component
	1
F9I1Y1	.494
F9I2Y1	.616
F9I3Y1	.370
F9I4Y1	.551
F9I5Y1	.390
F9I6PY1	.619

Extraction Method: Principal Component Analysis.
a 1 components extracted.

Form 9 Year 2

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	1.357	22.616	22.616	1.357	22.616	22.616
2	1.067	17.785	40.400	1.067	17.785	40.400
3	.936	15.600	56.000			
4	.919	15.309	71.309			
5	.904	15.068	86.377			
6	.817	13.623	100.000			

Extraction Method: Principal Component Analysis.

Component Matrix(a)

	Component	
	1	2
F9I1Y2	.444	-.509
F9I2Y2	.624	-.196
F9I3Y2	.318	.582
F9I4Y2	.505	.086
F9I5Y2	.307	.637
F9I6PY2	.566	-.134

Extraction Method: Principal Component Analysis.
a 2 components extracted.

Form 10 Year 1

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	1.419	28.373	28.373	1.419	28.373	28.373
2	.986	19.711	48.084			
3	.927	18.535	66.619			
4	.906	18.115	84.735			
5	.763	15.265	100.000			

Extraction Method: Principal Component Analysis.

Component Matrix(a)

	Component
	1
F10I1Y1	.460
F10I2Y1	.613
F10I3Y1	.444
F10I4Y1	.381
F10I5PY1	.698

Extraction Method: Principal Component Analysis.
a 1 components extracted.

Form 10 Year 2

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	1.387	27.735	27.735	1.387	27.735	27.735
2	.990	19.792	47.528			
3	.906	18.119	65.647			
4	.887	17.732	83.379			
5	.831	16.621	100.000			

Extraction Method: Principal Component Analysis.

Component Matrix(a)

	Component
	1
F10I1Y2	.502
F10I2Y2	.560
F10I3Y2	.418
F10I4Y2	.557
F10I5PY2	.579

Extraction Method: Principal Component Analysis.
a 1 components extracted.

Form 11 Year 1

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	1.691	28.176	28.176	1.691	28.176	28.176
2	.978	16.308	44.484			
3	.914	15.227	59.710			
4	.872	14.534	74.244			
5	.815	13.579	87.824			
6	.731	12.176	100.000			

Extraction Method: Principal Component Analysis.

Component Matrix(a)

	Component
	1
F11I1Y1	.431
F11I2Y1	.641
F11I3Y1	.429
F11I4Y1	.584
F11I5Y1	.514
F11I6PY1	.553

Extraction Method: Principal Component Analysis.
a 1 components extracted.

Form 11 Year 2

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	1.548	25.804	25.804	1.548	25.804	25.804
2	.985	16.413	42.217			
3	.942	15.705	57.922			
4	.891	14.846	72.769			
5	.851	14.183	86.951			
6	.783	13.049	100.000			

Extraction Method: Principal Component Analysis.

Component Matrix(a)

	Component
	1
F11I1Y2	.392
F11I2Y2	.649
F11I3Y2	.354
F11I4Y2	.577
F11I5Y2	.470
F11I6PY2	.542

Extraction Method: Principal Component Analysis.
a 1 components extracted.

Form 12 Year 1

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	1.552	31.044	31.044	1.552	31.044	31.044
2	.964	19.283	50.327			
3	.892	17.830	68.157			
4	.817	16.336	84.494			
5	.775	15.506	100.000			

Extraction Method: Principal Component Analysis.

Component Matrix(a)

	Component
	1
F12I1Y1	.662
F12I2Y1	.519
F12I3Y1	.472
F12I4Y1	.535
F12I5Y1	.580

Extraction Method: Principal Component Analysis.
a 1 components extracted.

Form 12 Year 2

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	1.561	31.224	31.224	1.561	31.224	31.224
2	.951	19.016	50.240			
3	.913	18.268	68.508			
4	.794	15.881	84.390			
5	.781	15.610	100.000			

Extraction Method: Principal Component Analysis.

Component Matrix(a)

	Component
	1
F12I1Y2	.646
F12I2Y2	.467
F12I3Y2	.440
F12I4Y2	.558
F12I5Y2	.648

Extraction Method: Principal Component Analysis.
a 1 components extracted.