9-2010

Item Parameter Drift as an Indication of Differential Opportunity to Learn: An Exploration of item Flagging Methods & Accurate Classification of Examinees

Tia M. Sukin
University of Massachusetts Amherst, tiacorliss@hotmail.com

ITEM PARAMETER DRIFT AS AN INDICATION OF DIFFERENTIAL OPPORTUNITY TO LEARN: AN EXPLORATION OF ITEM FLAGGING METHODS & ACCURATE CLASSIFICATION OF EXAMINEES

A Dissertation Presented

by

TIA M. SUKIN

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF EDUCATION

September 2010

School of Education
Educational Policy Research and Administration
Research Evaluation Methods Program

ITEM PARAMETER DRIFT AS AN INDICATION OF DIFFERENTIAL OPPORTUNITY TO LEARN: AN EXPLORATION OF ITEM FLAGGING METHODS & ACCURATE CLASSIFICATION OF EXAMINEES

A Dissertation Presented

by

TIA M. SUKIN

Approved as to style and content by:

____
Lisa A. Keller, Chair

____
Craig S. Wells, Member

____
George R. Milne, Member

____
Christine B. McCormick, Dean
School of Education

DEDICATION

To my strongest supporter, Michael Fechter,

for being patient, understanding, and full of rationale thoughts.

ACKNOWLEDGMENTS

My first thanksgiving must be extended to my dissertation committee chair, Lisa Keller. Her honest and dedicated guidance offered under the many hats of mentor, professor, colleague and friend will be forever remembered in my heart.

To follow, I thank Craig Wells and George Milne for agreeing to serve on my dissertation committee and providing their valued suggestions, advice, and feedback throughout this process.

For all the technical support and programming knowledge shared along with their offerings of good company, I thank my dear friends Rob Keller and Rob Cook. Without their forfeiture of a couple of extended weekends, the work contained within would remain unfinished for another five years, perhaps.

Much gratitude is expressed for my friends Su and Peter Baldwin for all their early counsel during my period of choosing whether to continue in the field of psychometrics; and especially thank you to Su for being such a wonderful and supportive friend for many years.

I thank distinguished professor Ronald Hambleton for his warm words of encouragement and for sharing his fantastic recognition of how the varied concepts of measurement weave together throughout its history.

Thank you to Steve Sireci for being so approachable and for encouraging me to apply to the Research Methods and Evaluation Program at this university during our conversation in New York at the annual conference for the National Council on Measurement in Education.

For providing formatting advice and genuine concern, I thank Peg Louraine. Her soft motherly care for the external lives of students that pass through her office are certainly appreciated.

I thank Michael Nering and Measured Progress for providing the funding necessary to complete this work. I also thank Rob Keller and Michael Nering at Measured Progress for fulfilling my data needs.

For ensuring that I spend a little time outside of my office, I thank my new found friend, Amy Semerjian. Her creativity, energy, and special drink mixes have brought much enjoyment to my life.

Finally, a wealth of appreciation is extended once again to Lisa Keller for motivating me to seek the help I needed. Thus, I especially thank my team of medical and holistic care professionals: Scott Belanger, Linda Cimma, Christine Horn, Jack Radner, Christopher Shanky, and Amy Winters who have all helped me restore and retain my balance of body, mind, and spirit.

ABSTRACT

ITEM PARAMETER DRIFT AS AN INDICATION OF DIFFERENTIAL OPPORTUNITY TO LEARN: AN EXPLORATION OF ITEM FLAGGING METHODS & ACCURATE CLASSIFICATION OF EXAMINEES

SEPTEMBER 2010

TIA M. SUKIN, A.A., COMMUNITY COLLEGE OF RHODE ISLAND

B.S., RHODE ISLAND COLLEGE

M.A., UNIVERSITY OF MARYLAND, COLLEGE PARK

Ed.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Lisa A. Keller

The presence of outlying anchor items is an issue faced by many testing agencies. The decision to retain or remove an item is a difficult one, especially when the content representation of the anchor set becomes questionable by item removal decisions. Additionally, the reason for the aberrancy is not always clear, and if the performance of the item has changed due to improvements in instruction, then removing the anchor item may not be appropriate and might produce misleading conclusions about the proficiency of the examinees. This study is conducted in two parts consisting of both a simulation and empirical data analysis. In these studies, the effect on examinee classification was investigated when the decision was made to remove or retain aberrant anchor items. Three methods of detection were explored; (1) delta plot, (2) IRT *b*-parameter plots, and (3) the RPU method. In the simulation study, degree of aberrancy was manipulated as well as the ability distribution of examinees and five aberrant item schemes were employed. In the empirical data analysis, archived statewide science achievement data that was suspected to possess differential opportunity to learn between administrations was re-analyzed using the various item parameter drift detection methods. The results for both the simulation and empirical data study provide support for eliminating the use of flagged items for linking assessments when a matrix-sampling design is used and a large number of items are used within that anchor. While neither the delta nor the IRT *b*-parameter plot methods produced results that would overwhelmingly support their use, it is recommended that both methods be employed in practice until further research is conducted for alternative methods, such as the RPU method since classification accuracy increases when such methods are employed and items are removed and most often, growth is not misrepresented by doing so.

TABLE OF CONTENTS

CHAPTER

APPENDICES

LIST OF TABLES

LIST OF FIGURES

Figure Page

CHAPTER 1

INTRODUCTION

1.1 Background

The accurate measurement of educational growth and progress has been essential in the era of No child Left Behind (NCLB[1], 2001) and will continue to be important as we move into the next reauthorization of the Elementary and Secondary Education Act (ESEA, 1965). However, it has been long lamented in the psychometric community that the measurement of growth is wrought with complications and may very well be one of the most difficult, yet consequential, tasks that psychometricians are faced with. Matters are further complicated when scores from different test forms administered in different years to different examinees are expected to be comparable. Many procedures are available for the scaling and equating of tests forms that will allow for score comparability. However, as states, districts, schools, and teachers adjust their instructional foci and curricula to meet the demands of higher standards, the traditional steps involved for equating test forms, mainly the retainment and removal decisions made for items that show signs of item parameter drift (IPD) in the non-equivalent groups anchor test (NEAT)[2] equating design, may mask or distort the true progress between cohorts of examinees.

The comparison of scores from different test forms through equating is made possible by the property of parameter invariance inherent in item response theory (IRT). Thus, item parameters are assumed to be invariant to the sample of examinees that

[1] For the reader's convenience, an alphabetized reference list of acronyms is provided in Table A.1 of Appendix A.

[2] Also frequently referred to as the common items non-equivalent groups (CINEG) design.

respond to items, and person parameters are assumed to be invariant to the set of items to which the examinee responds. While item parameters are invariant, they are only invariant up to a linear transformation, which results in the so-called identification problem (Hambleton, Swaminathan, & Rogers, 1991). The identification problem is usually resolved using one of the popular IRT scaling techniques: Mean-sigma (MS), mean-mean (MM), Stocking and Lord (SL), Haebara (HB), or fixed common item parameter (FCIP).

While the property of parameter invariance is a property of the parameters, it is often applied to the parameter *estimates*. To the extent that the estimates are not invariant, the scaling and equating that result from using this property may not be accurate. As such, it is required by the American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (1999) by way of *Standard* 4.13 that testing companies evaluate the functioning of the common (i.e., anchor) items used for equating as quoted below.

> *Standard* 4.13: In equating studies that employ an anchor set design, the characteristics of the anchor set and its similarity to the forms being equated should be presented, including both content specifications and empirically determined relationships among test scores. If anchor items are used, as in some IRT-based and classical equating, the representativeness and psychometric characteristics of anchor items should be presented.

Satisfying the later part of *Standard* 4.13 is usually done by comparing the parameter estimates of the common items on the two test forms; if the relationship between the item parameters is not linear, then the invariance of the estimate is suspect. The removal of aberrantly performing anchor items within a NEAT design from the equating of two or more forms is common practice. This practice is justified because the intent of the anchor items is to act as a set of items with parameters that are free from sample dependencies used to link forms (Yen & Fitzpatrick, 2006). However, in the presence of shifting instructional focus and curriculum, the likelihood of common items exhibiting item parameter drift (IPD) increases; and thus, the practice of item removal decisions may warrant further investigation so as to ensure the accurate measurement of gains in performance.

IPD can be thought of as a special case of differential item functioning (DIF) between test administrations. In studies of DIF, a focal group is defined and often that group is thought of as being potentially disadvantaged by the assessment (e.g., females or Native Americans) while the reference group is the set of examinees that the focal group is being compared to (e.g., males or Caucasians). The reference and focal groups can also be defined as ‘Administration 1’ and ‘Administration 2’ examinees, respectively where the first administration may be delivered one year and the second in another. Parameter estimates for the same items that vary across test administrations are therefore determined to possess IPD and thus those items function differentially between testing occasions. In light of the conceptual similarities between IPD and DIF, many of the methods employed to assess DIF within a test can be applied to the assessment of IPD across test forms when common items are used to equate them. Thus, it is important to clarify some often confused points regarding DIF and its detection.

1.1.1 DIF, Impact, and Bias --- What’s the Difference?

The difference in performance that may exist between intact groups is considered impact and differs from differential item functioning (Holland & Thayer, 1988) in that examinees within a group consistently perform higher or lower than another group on all items. Conversely, when examinees from the same ability groups (i.e., examinees are matched on some criterion, usually total test scores) but from different intact groups, perform differently on select items, this is considered to be differential item functioning. In the context of this work, differential item functioning is described as item parameter drift between administration years due to changes in curricular focus and emphasis.

Further, DIF, by proxy, indicates multidimensionality and thus represents a difference in a secondary ability or item parameters after conditioning on the skill or ability the test intended to measure (Camilli & Shepard, 1994; Roussos & Stout, 1996). DIF is a necessary but not a sufficient condition for bias. For bias to exist, the secondary ability must be an unintended component or not relevant to the purpose of testing. Therefore, bias can not be declared until after a content review has been conducted. In both cases, the difference that exists can not be attributed to random error of measurement. Thus, DIF is determined based upon item statistics alone; while, bias is determined only after follow up studies prompted by the DIF statistics.

Internal measures (i.e., comparing performance of individual items using total test scores as a conditioning variable) of DIF are only capable of detecting relative discrepancies and not constant bias (Camilli & Shepard, 1994) due to the artifact of conditioning on the total score of examinees when conducting DIF studies. Constant bias can be defined as all or many items that are equally measuring something unrelated to the content intended for measurement. For example, if the focal group is aware that they are being tested to evaluate the efficacy of the measurement and are not to be scored, they may collectively decide not to take the assessment seriously and thus may not try very hard on the test questions; thus producing the misleading results that they are, on the whole, not as capable as the reference group. Thus, the secondary trait measured would be motivation and this would not be considered relevant to the trait of interest, but would affect all item parameter estimates similarly, thus creating an appearance of impact rather than bias. Likewise, if both the focal and reference groups were informed that they would not be personally held accountable for the results of the assessment, both groups would have artificially low scores and this would be a matter of assessment validity which will not be discussed here, but the point is that the absence of DIF is not sufficient for assessing whether assessment results are valid for their intended uses.

To summarize, impact suggests that a "real" group difference in ability on the trait or skill of interest exists; while evidence of bias indicates an "artificial" group difference due to the measurement instrument itself (e.g., items flagged as differentially performing between groups). Therefore, DIF is declared when bias exists, but not in the presence of impact. Further, DIF is a necessary but not sufficient condition for bias.

1.1.2 Uniform vs. Non-uniform DIF

Uniform DIF exists when one group is consistently advantaged across the entire ability scale such that when looking at item characteristic curves (ICCs) for each of the groups, the curves do not cross. Therefore, uniform DIF exists when the *b*-parameter (i.e., difficulty) is the only parameter that is different between the groups. Conversely, non-uniform DIF exists when a group is advantaged or disadvantaged depending upon the point or interval within which the group members fall; thus, the ICCs cross. Therefore, non-uniform DIF exists when the *a*-parameter (i.e., discrimination) is different between groups. Non-uniform DIF is sometimes left undetected unless the procedure used to detect DIF is designed to specifically recognize non-uniform DIF. Camilli and Shepard

(1994) claim that non-uniform DIF is rare while Hambleton and Rogers (1989), for example, cite the detection of non-uniform DIF in practice for a high school proficiency exam. With regards to anchor items, it is quite plausible that items in the first year may be more discriminating and harder than they are in the second year of their administration. Thus, non-uniform DIF is expected to be relevant in the analysis of anchor items.

1.1.3 Effect Size vs. Statistical Significance

In terms of DIF, an effect size is the actual measurement of DIF that can provide a more practical level of importance; while statistical significance refers to the testing of the DIF statistic for significance given some null hypothesis (Camilli & Shepard, 1994; Holland & Thayer, 1988; Kim & Cohen, 1995). The latter is more sensitive to sample size, but may provide more clear-cut decision rules regarding item removal and inclusion decisions when selecting items for an anchor set. However, Longford, Holland, and Thayer (1993) suggest that despite the clarity that statistical tests seem to provide, effect sizes should be used, especially for comparing indices across time or testing occasions, mainly due to the sensitivity of the statistical tests to sample size; although this is not of great concern when the groups are of similar size from year to year as they usually are when comparison groups consist of examinees from different administration years.

1.2 Statement of Problem

When changes in achievement occur between test administrations, as would be predicted in assessments used for statewide accountability programs aimed at increasing the percent of students performing at a proficient level to 100%, caution must be exerted before removing anchor items that show a differential shift in difficulty or discrimination.

This shift could be due to changes in instruction and curricular emphasis, exposure of the item, differential printing of the item in the two forms, the presence of another item that provided clues to answering the item in question, to name a few reasons. In these instances, a decision must be made to include the item in the equating or not as the quality of the equating procedure applied relies strongly upon the adequacy of the common items used (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999; Yen & Fitzpatrick, 2006) and hence the accurate classification of examinees into proficiency groups.

While exposure, differential printing, and the presence of another item that provided clues to answering the item in question correctly are not desirable and would indicate a need to remove the offending item entirely; positive changes in performance due to a change in instruction is exactly what is sought in education, and removing an item for which this is true may be detrimental to the proper assessment of growth. Since it is not always clear which of these causes are responsible for the change in the performance of the item, it is essential to know what effect retaining or removing the item, either from the equating or both the scoring and equating would have on the classification of examinees into performance categories.

Further, research has shown that which items are flagged as differentially performing is, in part, dependent upon the method chosen for flagging items. Thus, it is important to choose reliable flagging methods and develop robust evaluation criteria for such methods.

1.3 Purpose of Study

There have been studies that have examined the effect of removing anchor items using empirical test data; however, there are few studies that have examined the effect of removing items from the equating on the accuracy of the classification of students into performance categories using simulated data. The benefit to using simulated data is that the true classification of the examinees is known, and the effect of retaining or deleting an anchor item on the accuracy of the classification of students can be ascertained. This study investigates the effect of removing outlying anchor items on the classification of students into performance categories. Additionally, it does so by using several different criteria for determining whether the item is an outlier. The importance of this study is clear in regards to the current large-scale educational assessment climate, where the accurate classification of students into performance categories is essential. Additionally, with assessments influencing the content of instruction, the likelihood of finding outlying anchor items is high and so deciding how to appropriately detect and deal with these items is of the utmost importance.

Thus, the purpose of this study is twofold and intends to provide answers to, (1) Which method(s) for flagging outlying anchor items perform(s) best, as determined by an analysis of Type I error rates and power; and (2) Does removing items from the anchor significantly mask the magnitude of growth as assessed by proficiency classification accuracy? Each of these questions will be answered via simulation data analyses informed by empirical assessment item parameters. Following the simulation studies, the detection methods will be applied to empirical data. Consequences arising as a result of differences or the lack there of between current practice and simulation study conditions will be addressed. Ultimately, this study is intended to inform large-scale standardized achievement assessment equating practices.

What follows is an extensive literature review of both equating using the NEAT design and a comparison of DIF detection methods and their applications to the identification of aberrant anchor items. Following the literature review is a thorough outline of the methods used for this study, including descriptions of the models, technical details of the DIF detection methods employed, and all study conditions. Next, the results of the simulation and empirical data study are presented along with comparisons to current practice. Finally, a discussion of the implications and educational importance of these results is presented.

CHAPTER 2

REVIEW OF LITERATURE

2.1 Overview of Literature Review

This chapter reviews the literature related to the NEAT equating design along with IPD detection procedures. More specifically, this chapter can be outlined into the following six sections:

1. NEAT Equating Design. This section focuses on a description of the NEAT equating design along with its advantages over the single-group and equivalent-group designs. Further, extensions of the NEAT equating design to matrix-sampling are explored.
2. Anchor Set Composition. This section reviews the multitude of studies conducted to provide recommendations on constructing an anchor set (i.e., the set of items used for linking assessment forms). Major components include; (1) length, (2) content representation, (3) statistical properties, (4) item parameter drift, and (5) utility and placement of the anchor set and linking items within it.
3. Population Invariance & Opportunity to Learn. This section specifically reviews studies that explore opportunity to learn (OTL) as a contributing factor to observed population invariance.
4. Scaling. This section provides a short summary of studies conducted that compare the performance of scaling methods as applied to the accurate measurement of growth.

5. Identification of Aberrant Items. This section reviews the existing research that uses both item parameter drift and differential item functioning detection methods for the identification of aberrant items.
6. Impact on Ability Parameter Estimation & Classification. This section presents studies performed that assess the impact of item removal and retainment decisions of a linking item within an anchor set on the estimation of ability parameters and the classification of examinees.

This chapter concludes with a summary of the literature reviewed.

2.2 NEAT Equating Design

Educational test practitioners realize the need for ensuring a quality process for equating two or more forms of a test. Typically, when referring to criterion-referenced state-achievement exams, examinees who take a particular test form are considered to be part of a naturally occurring group. This means that the examinees are not randomly selected from some specified population, but instead, groups of examinees are formed based on the requirements of testing. Statewide testing that occurs in grades 3-12 for accountability purposes are examples. It is unlikely that examinees taking the exam from one year to the next are equivalent; thus, ruling out use of the randomly equivalent groups design. This is especially true with the desire for schools to meet Annual Yearly Progress (AYP). Therefore, it is likely that the group of examinees taking later forms of the test have higher mean ability due to curricular and instructional modifications intended to better prepare students on all academic standards tested. Non-equivalent groups make equating test forms difficult as test forms are rarely strictly parallel resulting in one form being slightly more difficult than another. Differences between examinee groups and test forms must be accounted for when equating test forms. The most common test equating design used to adjust for these differences is the NEAT design. This well-known design consists of using a set of items that appear on each of the tests to be equated and are often called anchor or common items. In the context of classical test theory (CTT), the anchor items are used to adjust for ability differences in the naturally occurring groups of examinees (e.g., Angoff, 1968; Angoff, 1971; Gulliksen, 1950; Holland & Dorans, 2006; Kolen & Brennan, 2004; Petersen, Kolen, & Hoover, 1989). Within the context of item response theory (IRT), the anchor set is used to place item parameter estimates onto the same scale (e.g., Hambleton, Swaminathan, & Rogers, 1991; Holland & Dorans, 2006; Kolen & Brennan, 2004; Lord, 1980). Scores can then be adjusted for the differences in the difficulties of the two tests, thereby preserving the differences due to differing ability (Brennan, 2006; Lord, 1980). Additionally, the NEAT design is the most practical for most testing situations and avoids problems associated with the single-group (e.g., effects of practice or fatigue, cost, and practicality) and equivalent-group (e.g., often, groups are not equivalent) designs when anchor items used represent the content and item specifications of the full test (Hambleton, Swaminathan, & Rogers, 1991).

The NEAT method can be further extended to equate tests over several test forms using the matrix-sampling equating design as shown in Figure 2.1. Anchor items are spread across forms such that not all examinees respond to all anchor items, but instead; given large samples, item parameters for common items are obtained from a random selection of individuals. By design, the anchor set is external which means that the items contained within are not used for scoring examinees. This design is practical when large numbers of examinees are assessed each year, thus making it possible to scale test forms based on several anchor items, but reducing the number of extra items each examinee must complete. Figure 2.1 provides the conceptual structure of such a design. Items within the common form[3] are administered to all examinees and are used for scoring. These are the only items that are used for scoring and the common form of items will change from year to year. The matrix-sampled items appear on the lettered forms (i.e., Form A, Form B, and Form C in the example presented) and differ across forms. However, the items within these forms are used for deriving the scaling constants to place each of the test forms onto the same scale as these items would have appeared on the previous year's test. Thus, Form A matrix items would be the same items from year to year and would be used for placing assessments on the same scale. The same would be true for Forms B and C in the example presented here.

2.3 Anchor Set Composition

Five important features of an anchor set are length, content, statistical properties, invariance over time (i.e., lack of item parameter drift), and utility/placement of the anchor set and common items. Each of these features will be described in detail with the inclusion of recommendations made for forming the anchor set based upon a review of the literature.

2.3.1 Anchor Set Length

It is well known that the length of a single test is highly correlated with its reliability. Longer tests are often more reliable than shorter tests that measure the same construct (Angoff, 1968). Therefore, it would also be reasonable to assume that the length of an anchor set should be such that its reliability is of a respectable level; as high reliability is a necessary condition for valid interpretations of test scores. This is

[3] Not to be confused with the term, "common items" which make up the anchor set.

exemplified if we consider a student that takes a test on one occasion and again on another occasion (i.e., assuming nothing about the person's trait being measured has changed) receiving two distinct scores based on a lack of reliability. Given these different scores, different conclusions would be drawn for that student, thus indicating that claims of validity for the interpretations based on either of the distinct scores would be severely suspect. Thus, it is clear when discussions of validity arise, a case for reliability must also be given. Most of the research suggests that the anchor set should represent at least 20% of the operational test; or for IRT equating methods, at least 15 items should be used to properly serve this purpose (e.g., Angoff, 1968, 1971; Brennan & Kolen, 1987; Cook & Eignor, 1991; Fitzpatrick, 2008; Hambleton, Swaminathan, & Rogers, 1991; Kolen & Brennan, 2004; McKinley & Reckase, 1981; Vale, Maurelli, Gialluca, Weiss, & Ree, 1981; Wingersky, Cook, & Eignor, 1987). Studies such as Keller, Egan, and Schneider (2010) even suggest upwards of including 25% of the operational test as linking items, when using an internal anchor, as one of their important findings in varying the number of items used for equating (i.e., 6, 11, and 14) resulted in the ability to appropriately detect items as aberrant as the anchor set became sufficiently long.

2.3.2 Content Representation

While the operational test may meet all the requirements for the valid interpretation of test scores, including content relevancy, the anchor set may violate content representativeness if the balance of test items measuring different test specifications is not maintained within the anchor. Therefore, particular content areas may become over- or under- valued in the linking of test forms especially in the presence of multidimensionality. While the operational test may be "essentially unidimensional," it is likely that some multidimensionality exists and this can be accentuated in the improper balance of content in the anchor. Furthermore, all of the relevant and available documentation speaks to the importance of creating an anchor set that proportionally follows the content specifications of the operational test when using the NEAT design (e.g., Cook, & Eignor, 1991; Cook & Petersen, 1987; Dorans, Kubiak, & Melican, 1998; Hambleton, Swaminathan, & Rogers, 1991; Klein, & Jarjoura, 1985; Kolen, 1988; Kolen, 2004; Kolen, & Brennan, 2004; Petersen, Kolen, & Hoover, 1989; Petersen, Marco, & Stewart, 1982; Sinharay, & Holland, 2006, 2007, 2008; Yang, 2000).

Klein and Jarjoura (1985) compared a content representative anchor to a long anchor without content representation and found that the shorter anchor with content representation proved to perform best when using Tucker linear equating and Levine equating. This is the study often cited supporting the recommendation that anchor and operational tests contain equivalent proportions of items representing differing content areas. Very few sources were found that confirmed these findings using IRT methods for equating, which usually assume unidimensionality of the construct. Yang (2000) studied four anchor item sampling designs and four equating methods, two of which utilized IRT designs. Under all methods of equating, Yang (2000) found that equating accuracy was greatest when using the item-sampling scheme that selected items for inclusion in the anchor set in such a way that the anchor items proportionally matched the content specifications for the entire test. However, the lengths of the total tests and the anchor sets were not the same across all four subtests, thus confounding the results. Again, these last two studies speak to the nature of “essential unidimensionality” such that as long as content proportions are maintained between operational tests and within the anchor set, analyses and scoring can take place under the assumptions of unidimensionality, thus potentially being the important reason that content specifications be maintained. However, two preliminary research studies (Sukin, Dunn, Kim, & Keller, 2010; Sukin & Keller, 2009a) exist that compare strict proportional sampling to optimal sampling designs that take the subtest variability into account when selecting items for inclusion on the anchor set. These studies both found that the optimal sampling designs employed may lead to satisfactory equating results as well. Regardless, in the event that unidimensionality exists in its purist sense, proportioning of content would be an arbitrary point. To the extent that the set of anchor items do not address each of the content areas proportionately, construct validity may be violated.

2.3.3 Statistical Equivalence to the Total Test

Similar to the research related to content matching, recommendations from the literature supports that the anchor set be composed of items that mimic the statistical properties of the operational test (e.g., Angoff, 1968; Cook, & Eignor, 1991; Dorans, Kubiak, & Melican, 1998; Kolen, 2004; Kolen, 1988; Kolen, & Brennan, 2004; Petersen, Marco, & Stewart, 1982; Petersen, Kolen, & Hoover, 1989; Petersen, Marco, & Stewart, 1982). Often, the anchor set is referred to as a “mini-test” and consists of items with similar mean difficulty and similar range of difficulty. As an example, Petersen, Marco, and Stewart (1982) studied several equating methods (i.e., external vs. internal, content similarity, mean difficulty similarity) with a total of 11 conditions using the Scholastic Aptitude Test (SAT) and the Test of Standard Written English (TSWE) and found that matching the mean difficulty of test and anchor items was a more important factor in ensuring a reliable anchor set for equating test forms when compared to the scenario where moderate differences in the content of anchor and operational tests existed. This conclusion was based on equating a test to itself via equipercentile methods. However, other research (Sinharay & Holland, 2006, 2007, 2008) has shown that preserving the standard deviation of the difficulty parameter from the operational test within the anchor set may not be necessary.

If one regards the operational test as the measurement device and the anchor set as the criterion, as in concurrent validity when two measures occur within the same timeframe, work of the early measurement specialists can be used to make an argument for valuing the correlation of anchor and operational test scores (e.g., Cronbach & Warrington, 1952). While Sinharay and Holland (2006, 2007, and 2008) support the "mini-test" configuration when using an internal anchor set design, they do not fully support this idea when using an external anchor. These authors present evidence supporting the "semi-midi" and "midi-test" forms as anchors instead of the "mini-test" form where the spread of the item difficulties are more constrained in order to preserve items that are either very easy or very hard in terms of the item's estimated difficulty parameter. Preserving these items protects against overexposure of items that may not be produced in large numbers, thus saving valuable time and resources. However, such preservation methods are not needed when the anchor items are also used for arriving at scoring decisions. These authors found that such anchor sets perform equally well and sometimes better than the "mini-test" when using poststratification and equipercentile equating methods. Additionally, the midi- and semi-midi anchor sets were found to have higher anchor-test-to-total-test correlations than the "mini-test." Further work must be conducted using IRT equating methods to generalize these results and determine whether value should be placed more so on the correlation of test scores, an indication of concurrent validity, or on the construct representation based on the item specifications within the test blueprint, an indication of content validity.

2.3.4 Item Parameter Drift

Statistical equivalence was discussed in terms of the anchor set's relation to the total test. Now we turn our attention to the stability of the statistical properties or parameters of the items used in the anchor set over time (i.e., across administrations). The process of assessing whether item parameter drift (IPD) exists for any given linking item between multiple test administrations can be related to differential item functioning (DIF) analyses where the focal and reference group each represent a testing occasion (e.g., the reference group is considered Administration or Year 1's examinees and the focal group is considered Administration or Year 2's examinees). Ackerman (1992) clearly explains that when DIF is observed, the DIF may be a result of item bias. However, DIF is not a necessary condition for item bias. Item bias is concluded to exist when the item is measuring something other than it purports to measure as discussed in the introduction. In this case, construct validity is violated due to the existence of construct irrelevant variance. When items are exposed, systematic cheating occurs, or the presence of surrounding items provide clues for a correct response for one of the linking items for one administration of the test and not the other(s), IPD may be observed. However, the observed DIF may not be considered item bias if the item is truly measuring the construct of interest. In this case, the observed DIF may be due to factors associated with curricular or instructional changes between test administrations that highlight the content or concept of the item in question (Bock, Muraki, & Pfeiffenberger, 1988), thus making the probability of a correct response for the item higher. Conversely, curricular or instructional changes can occur that would make the probability of a correct response lower, thus indicating that the item is harder. Therefore, concluding that all linking items that exhibit IPD are ‘bad’ items is not sound measurement practice. Likewise, it is not sound measurement practice to disregard IPD either. Without careful analyses, concluding that items that experience IPD are ‘good’ or ‘bad’ lead to the confounding of the measurement of growth for the construct of interest. In both cases, an over- or under-estimate of growth may occur that may lead to erroneous scaling of test forms and thus the false conclusion that scores on alternate test forms are comparable.

Further, it should be addressed here that while the example given for the reference and focal groups are ‘Year 1’ and ‘Year 2’ examinees, IPD can compound over several test administrations (Bock, Muraki, & Pfeiffenberger, 1988) such that IPD would not be adequately detected until the third or fourth administration, for example.

2.3.5 Utility & Placement of the Anchor Set

Anchor sets may either be considered internal or external and both consist of items common to the operational forms being equated. These items are used in the final scoring of examinees when an internal anchor set is employed while they are not used for scoring when an external anchor set is employed. For the latter, often the items are presented in a separately timed section of the test. However, they may also be scattered throughout the test as when using a matrix-sampling design. These methods each have their own unique advantages and disadvantages.

Extra testing time is needed when an external anchor set is used. This is because the items are not used in calculating scores for examinees and therefore, these items are considered extra items. Additionally, these items can be hard to disguise, except in the case of the matrix-sampling design; thus, they are also prone to construct irrelevant variance. Construct irrelevant variance can creep into test scores when traits or factors other than those intended for measurement are indeed measured. Examples include, motivation, fatigue, and other psychological influences associated with the knowledge that low-stakes are attached to performance on external anchor items. This becomes less of a concern when external items are embedded within the operational test or when examinees are unaware of which items are scored and which are not. However, even within the matrix-sampling design, issues related to validity exist. For example, when items are placed in different locations between test forms, which is often done within this design, contextual effects threaten the validity of the decisions made based on the assumption of score comparability (Eignor, 1985; Kingston & Dorans, 1984).

Internal anchor items are additionally subject to breaches of security because they are seen at a higher frequency by more examinees. The over-exposure of anchor items may lead to false estimates of the magnitude of observed growth from one test administration to the next if the exposure goes undetected. Despite these challenges, Marco, Petersen, and Stewart (1983) found that overall equating error was less for operational tests that employed internal rather than external anchor sets. However, these results are confounded by the fact that internal anchor sets more closely represented the statistical properties of the entire test than did the external anchor items.

Recommendations from the literature suggest that anchor items, whether used externally or internally, be placed in the same locations between test forms, and preserve wording and context (Angoff, 1968; Brennan, & Kolen, 1987; Cook & Petersen, 1987;

Kolen & Brennan, 2004; Yen, 1980). Cook and Eignor (1989), Kingston and Dorans (1984), and Whitely and Dawis (1976) all found that the susceptibility of anchor items to contextual effects depends on the item type and how many other items of the same type have already been seen, suggesting that ‘practice’ is sometimes measured as construct irrelevant variance. Conversely, it is not wise to place all anchor items at the end of the test as examinees may not reach all items due to time constraints (Angoff, 1968).

2.4 Population Invariance & Opportunity to Learn

Kolen (2004) provides an extensive summary of the history of and the work done to address population invariance in equating. He presents the work of Cook, Eignor, and Taft (1988) as an example. They studied population invariance by using a NEAT design to equate forms of a high school biology exam. One form equating was conducted after the administration of the exam in the fall of the students’ senior year (i.e., two years after taking the course) and the other set of form equating was conducted after the administration of the exam in the spring of the students’ sophomore year (i.e., upon completion of the course). Additionally, they investigated the use of different common item sets. Both studies affected the equating suggesting that, barring any error due to sampling, recency of instruction and content interacted to produce different equating results. Several groups of anchor items were assessed and the authors found that when the test measured the same underlying concepts for the two groups, choice of the set of common items used for the equating did not matter. This was not the case when comparing the performance of common items across the spring and fall administration groups. Instead, the choice of common items did matter. The authors point out implications of these results must be considered in terms of validity – the inferences made based on the test scores. “Should the test be designed to measure immediate end-of-course outcomes or perhaps more enduring concepts” (p. 44)? The conclusion of this study can be summarized as curriculum-related achievement tests have differential validity depending on when a student takes the test. Likewise, it is easy to extend this to say that if teachers are changing what and how curriculum is taught from year to year, an interaction may exist between the year of instruction and the construct being measured, thus making the comparison of test scores from one test administration to the next unadvisable. This work appeared to be the most relevant empirical research for supporting the need to address the effects of equating forms using the NEAT design when examinees taking the different forms have been exposed to differential instructional emphases, possibly due to enhanced preparation for exams to meet AYP goals.

Similarly, Miller and Linn (1988) looked at the effect of variations in instructional coverage on item characteristic functions. To do so, they examined the results from a mathematics achievement test and an opportunity to learn (OTL) questionnaire, completed by teachers. Curriculum clusters were formed by factor analyzing OTL questionnaire responses and then subjecting the factor loadings to a cluster analysis that revealed three curriculum clusters for each subtest (i.e., arithmetic & algebra). Item response curves were compared between the curriculum clusters using two measures of item invariance (i.e., unsigned sum of squared differences & signed sum of squared differences). Findings showed that the magnitude of difference between the item characteristic curves was larger for curricular groups than for those reported when comparing black and white students suggesting that OTL is a big factor in group achievement differences and that it may be more appropriate to state that “instructional bias" and not necessarily "item bias" explains these results, as suggested by Linn and Harnisch (1981). Miller and Linn (1988) conclude their report by drawing attention to the notion that opportunity to learn may not be the only factor associated with differential item functioning. Intensity and effectiveness of instruction along with the student's motivation to learn and the teacher's motivation to teach may also interact with the lack of item characteristic function invariance, providing evidence once again that systematic construct irrelevant variance is often measured, partly, in place of or in addition to the desired construct and that item parameter drift among the anchor items may be picking up on this phenomenon. To summarize, Wang (2004) makes the poignant point that, "Construct invariance over time is the prerequisite of change measurement." If the construct changes between measurement occasions, growth can not be assessed adequately. Yet, this is exactly what applied psychometricians are being asked to provide. Therefore, the field of psychometrics must explore how robust current methodologies are and move forward with refining methods toward this end.

2.5 Scaling

The literature has shown that each of the scaling methods; mean-mean (MM), mean-sigma (MS), Haebara (HB), Stocking and Lord (SL), and fixed-common item parameter (FCIP), results in slightly different outcomes; with some methods proving to be more robust than others to different testing contexts (Kolen & Brennan, 2004). For example, Hanson and Béguin (2002) found that the test characteristic curve methods (i.e., HB & SL) produced the best estimates of ability within the context of NEAT equating designs.

Upon the inspection of growth recovery, Jodoin, Keller, and Swaminathan (2003) found that the choice of which method to employ for equating test forms with a NEAT framework mattered. In this study, three years of state achievement data in mathematics was used to compare the amount of growth calculated between cohorts using three different methods for scaling test forms; linear, fixed common item, and concurrent parameter estimation. Each of the methods was compared based on the final ability parameter estimates on the placement of examinees into classification categories. In using the same high-stakes assessment data, findings showed that the fixed common item parameter and concurrent methods produced a larger amount of growth than the linear equating procedures. Additionally, the authors state, "To the extent that the chosen equating method is not appropriate, comparisons between equated scores may not be appropriate, either" (p. 3). Therefore, the interpretation of gain scores and changes in the percentage of students being classified as proficient across cohorts is suspect and thus findings resulting from assessments of school effectiveness may also be invalid. These findings suggest that mixing equating methods across years is likely to severely confound measures of growth as well (Jodoin, Keller, & Swaminathan, 2003).

Further, Skorupski, Jodoin, Keller, and Swaminathan (2003) found similar results using simulated data. Simulated data was used in this research in order to distinguish which methods provided the most accurate estimates of growth between test administrations. Several factors were manipulated in order to assess which methods perform best at recovering true growth under varied conditions. Factors manipulated for investigation included; method (i.e., concurrent calibration assuming equivalent groups, concurrent calibration assuming non-equivalent groups, fixed common item parameters

(FCIP), mean-sigma (MS), and test characteristic curve), ability estimation method (i.e., maximum likelihood, maximum a posteriori, and expected a posteriori), number of examinees per form (i.e., 500, 1,000, & 2,000), mean true growth in ability on the theta scale (i.e., 0, .25, & .50), and test length (i.e., 35, 40, 55, & 60 items). Results indicated that the five methods for equating under investigation did produce systematically different results. For instance, calibration techniques tended to underestimate growth while transformation techniques slightly overestimated growth. However, when the number of linking items was increased, calibration techniques performed better. Again, such discrepancies have important policy implications for the assessment of growth over time.

However, work that has focused on common item inclusion for anchor sets that have compared these scaling techniques have found that choice of technique does not lead to results that significantly differ when results are assessed based upon decision consistency (i.e., the placement of examinees into the same proficiency categories over conditions), holding all other study conditions constant (Sukin & Keller, 2008, 2009b). Concurrent scaling methods were not explored as assessing the anchor items would be impossible using such scaling methods.

2.6 Identification of Aberrant Items

This section provides a review of the literature that explores research regarding the detection of aberrant items. Where appropriate, research within the field of differential item functioning is also addressed.

IPD and DIF detection methods can be categorized as either empirical-based or model-based. Empirical methods include delta plots (Holland & Thayer, 1985) and

Mantel-Haenszel (Dorans & Holland, 1993) procedures. The model-based methods include plots of IRT parameter estimates (Huff & Hambleton, 2001; Kolen & Brennan, 2004), Lord's (1980) χ^2, Stocking and Lord's (1983) test characteristic curve (TCC) inverse, Raju's (1988) area measures, the likelihood ratio (LR) test (Thissen, Steinberg, & Wainer, 1993), differential functioning of items and tests (DFIT; Raju, van der Linden, & Fleer, 1995) and the RPU method (Keller, Egan, & Schneider, 2010). With each of these methods, in order to obtain trustworthy results, the model must fit the data and a sufficient sample size must be used.

Use of delta plots, like many other detection methods, were first primarily used and developed for the detection of DIF. For instance, research surrounding DIF detection methods involving internal methods developed in the late 1970's showed that such methods as the transformed item difficulty[4] (TID, Angoff & Ford, 1973), chi-square measures developed by Scheuneman (1979) and full chi-square (Camilli, 1979), and area IRT methods of Rudner (1977) were all correlated highly with one another (Merz & Gossen, 1979; Rudner, Getson, & Knight, 1980). Further, Ironson and Subkoviak (1979) and Shepard, Camilli, and Averill (1981) found that the full chi-square seemed reasonable as an alternative to the ICC methods when sample sizes were small. Signed indices correlated more highly as expected since the chi-square methods are often not sensitive to non-uniform DIF. These early findings are synonymous with the more current findings using the more refined methods as will be seen in the review of the literature that follows.

The sections that follow are organized by a presentation of convergent and divergent research studies regarding the ability of methods to control Type I error rates

[4] Synonymous with the delta plot method.

for item detection, power, ability to detect uniform and non-uniform DIF, availability of an effect size measure or statistical significance test, and complexity and practicality. For each of these five criteria for comparing methods, dichotomous and polytomous methods along with empirical-based and model-based methods are explored.

2.6.1 Type I Error Rate

Type I errors occur when items are falsely identified as DIF. This often occurs when methods are sensitive to sample size or the discriminating power of the item under study. For the delta plot method (a.k.a. TID), Type I errors are an artificial appearance of DIF that occurs when items are highly discriminating (i.e., based on point biserial correlations) and the item appears between the group means on the ability scale due to the difficulty and discrimination interaction. Thus, typically, the more discriminating items will show more DIF between groups than less discriminating ones; such that highly discriminating items, often perceived in item reviews as good items, are frequently miss-flagged during DIF investigations (Camilli & Shepard, 1994). Therefore, the delta plot method has been declared as unsuitable for studying item bias, regardless of how *p* (i.e., the difficulty index) is transformed (Lord, 1980).

Other methods such as the Mantel-Haenszel (MH) procedure tend to have low Type I error rates so long as the groups being compared have equal mean abilities and the DIF is uniform (Gierl, Jodoin, & Ackerman, 2000; Kristjansson, Aylesworth, McDowell, & Zumbo, 2005; Narayanan & Swaminathan, 1996; Penfield & Lam, 2000; Roussos & Stout, 1996). However, the Type I error rate increases as groups differ in mean ability, discrimination of the item increases, and the reliability of the matching variable decreases

(Penfield & Lam, 2000; Su & Wang, 2005; Zwick & Thayer, 1996; Zwick, Thayer, & Mazzeo, 1997).

Model-based approaches to DIF detection are sensitive to sample size. For instance, Raju (1990) developed statistical tests of significance for his area measures and found that they worked well for the 1- and 2- parameter models, but did not work well under a fully estimated three-parameter model. Later, Camilli and Shepard (1994) made the recommendation that Lord's (1980) chi-square not be used as false rejection of the null hypothesis is common place suggesting that the measure may be too sensitive. Thissen and Wainer (1982) maintain the belief that this false rejection rate may be due in part to the inaccuracy involved in the estimation of the variance-covariance matrix required. Thus, when a different combination of discrimination, difficulty, and pseudo-guessing parameters are produced, yet similar ICCs exist such that DIF would not be present, the Lord chi-square procedure detects DIF (Camilli, 2006). In the empirical analyses of Kim and Cohen (1995), this did not present as a problem. However, their analyses dealt with ideal conditions where the comparison groups had similar ability distributions. The likelihood ratio test relieves the concern of false rejections due to inaccurate estimation of the covariance matrix by testing the entire item response function (Camilli, 2006). Further, Kim and Cohen (1997) presented that when the likelihood ratio test is used for DIF detection using the general response model, Type I error rates under varying sample sizes (i.e., 300 and 1,000) and ability matching conditions were consistent with their expected values.

While some disagreement occurs, the general consensus that all methods struggle with Type I error rates when using statistical tests of significance suggests the need for effect size measures. One such method that relies upon effect size is the RPU method (Keller, Egan, & Schneider, 2010) and will be discussed later in Section 2.6.4.

2.6.2 Power

Power is the ability to detect items that are functioning differentially between groups. Thus, if an item is declared as non-DIF and it truly is, the power of the detection method is weak. In using the delta plot method, real DIF may be missed when items are especially easy or especially hard (Camilli & Shepard, 1994). Therefore, results are often confounded with true group differences (i.e., impact). Other empirical-based methods (e.g., MH) tend to have high power in the detection of items with DIF with a tendency for the power to decrease as more non-uniform DIF items exist (Kristjansson, Aylesworth, McDowell, & Zumbo, 2005; Penfield & Lam, 2000).

Several comparison studies exploring IPD detection using the delta plot method have consistency revealed its inadequacies due to lack of power within varied contexts. Michaelides (2006) found the MH method to outperform the delta plot method for identifying aberrant anchor items across four assessments. Karkee and Choi (2005) found that while IRT-based measures flagged similar items for IPD, the delta plot method did not flag any items. Further, Sukin and Keller (2009b) examined the delta plot method in comparison to the TCC inverse and likelihood ratio test methods under simulated conditions of 0.5 and 0.8 drift for *b*-parameter values for selected items and with ability distributions of N(0,1) and N(1,1) and found the delta plot method to have zero detection power under all conditions. Likewise, studies conducted by Keller, Egan and Schneider (2010) and Sukin and Keller (2010) varied the aberrant item schemes (i.e., number of items simulated as aberrant with differing locations along the item difficulty scale) and found similar poor power for the delta plot method.

Among model-based methods (i.e., Lord's chi-square, Raju's area measures, and the likelihood ratio test) Kim and Cohen (1995) cite close agreement especially between methods when using iterative procedures for DIF detection. Similarly, Donoghue and Isham (1998) found that Lord's chi-square (i.e., when the *c*-parameter was constrained to be equal across assessment years) and the *z*-statistic for the unsigned area measure to be the most effective for DIF detection even in the presence of *ab*-drift and under bi-directional drift. However, the likelihood ratio test was found to lack power with small sample sizes (e.g., n=500); but, it was found to be more powerful when ability distributions differed between groups when compared in a simulation study to the Mantel chi-square method (Ankenmann, Witt, & Dunbar, 1999). Further, the likelihood ratio test was found to perform with perfect power when preliminarily (i.e., a total of five replications per study condition were implemented) investigated by Sukin and Keller (2009b). Conversely, findings showed the TCC inverse method to be inconsistent as did the findings presented in the Keller, Egan, and Schneider (2010) study. The latter study additionally found area measures to be quite promising especially when operationalized as the RPU method where several statistics of area are computed and compared to predetermined effect sizes.

Generally, power seems to be mostly a concern for the empirical-based methods as model-based methods tend to flag more items, either correctly or incorrectly (i.e., Type I error). Power seems to become more of a concern when items are characterized as possessing non-unifiorm DIF as discussed in the next section.

2.6.3 Uniform & Non-uniform DIF

As described in the introduction, uniform DIF occurs when the probability of responding correctly to an item is affected in the same way across the entire ability continuum of the focal group; and non-uniform DIF occurs when the probability of responding correctly to an item is affected in differing ways across the ability continuum depending on the examinees' location on that continuum. Therefore, the ICCs for the reference and focal groups cross. This kind of DIF is more difficult to detect as some methods do not account for the canceling out effect that occurs between positive and negative regions of area. The delta plot method would typically not detect items with non-uniform DIF as the overall probability differences for obtaining a correct response between groups may not differ if the ICCs cross in the middle of the ability continuum. Other empirical-based methods (e.g., MH) also often miss items with non-uniform DIF except for the logistic regression methods which are designed for the detection of both uniform and non-uniform DIF (Hambleton & Rogers, 1989; Hills, 1989; Narayanan & Swaminathan, 1996; Penfield & Lam, 2000; Spray & Miller, 1994; Swaminathan & Rogers, 1990; Zumbo, 1999).

Further exceptions do exist as Mazor, Clauser, and Hambleton (1994) applied a simple modification to the MH method and found that this modification was moderately successful in detecting non-uniform DIF. The modification involved splitting the examinee group at the middle of the ability distribution and analyzing each group separately for DIF using the traditional MH method. The rationale was that since the MH method most often fails at detecting non-uniform DIF when the crossing of the ICCs occurs in the middle of the distribution and for items of medium difficulty (Rogers, 1989) that the separate analyses would prevent the canceling effect that occurs for the positive and negative DIF observed across the entire scale.

Of the model-based procedures, the signed area measure is incapable of detecting non-uniform DIF by design; while the unsigned area measure is designed to detect non-uniform DIF, but at the cost of distorting the magnitude due to sampling error (Camilli & Shepard, 1994). Thus, the RPU procedure incorporates both types of statistics for making final flagging decisions, but no published research currently exits to support the efficacy of this method for flagging items for uniform and non-uniform drift. Further, the DFIT framework incorporates the analysis of items via non-compensatory (NCDIF) and compensatory (CDIF) measures such that NCDIF measures contain the ability to detect both uniform and non-uniform DIF while CDIF does not detect non-uniform DIF (Oshima & Morris, 2008).

Overall, the results of the ability of measures to detect non-uniform DIF are consistent and seems to favor the likelihood ratio test and DFIT framework as the best methods for the detection of both uniform and non-uniform DIF of the model-based procedures and logistic regression for an empirical-based method.

2.6.4 Effect Size Measures & Tests of Significance

As seen in the literature reviewed regarding the common inflation of Type I error rates when statistical tests of significance are conducted leads to the support for effect size measures that take the standard error and sample size into account. The advantage of the MH procedure is that both statistical tests of significance and effect sizes are used in conjunction with one another to categorize the severity of DIF that exists within a practical framework.

Another interesting approach that makes use of effect sizes, capitalizes on the various ways of describing the difference between ICCs. This approach, the RPU method, calculates seven difference statistics for all anchor items and then the item in question is flagged if four or more of the statistics exceed predetermined effect size thresholds. After reviewing a host of item detection methods, Keller, Egan, and Schneider (2010) found this method to be quite effective within the context of an internal anchor for correctly flagging aberrantly performing items.

The strength of the MH approaches to DIF detection is that well-established effect size and significance criteria have been set and tested (i.e., developed by ETS for dichotomously scored items and by NAEP for polytomously scored items). The RPU method does not have as rich of a published history as the MH procedure, but has been shown to be promising.

2.6.5 Complexity & Practicality

As with anything that needs to be implemented in practice, the assessment of the complexity and practicality of a method must be taken into account. This section explores this factor given the advantages and disadvantages of the methods discussed thus far.

Computationally, the empirical-based methods (i.e., delta plots and MH) are simpler than the model-based methods (Camilli & Shepard, 1994; Penfield, 2001). So long as a matrix-sampling design is not employed, MH methods are sometimes preferred as they are practically more convenient because they are less time consuming to conduct and less costly than some of the model-based procedures (Camilli & Shepard, 1994). However, MH methods can only be used if a sufficient number of items are common between forms. In the case of the matrix-sampling design, common items are often spread across many forms and the items used to form the matching variable becomes limited, thus ruling out the practical use of this method under the matrix-sampling equating design (Michaelides, 2006, 2008).

For all the model-based methods for DIF assessment, large samples are required, the data must fit the model, and users must possess a practical knowledge of item response theory. Typically, the problem associated with the area measures is that the differences in item response functions are distorted when individuals are sparsely located at the extremes of the θ continuum and this is where the differences occur. Also, if *c*-parameters differ, integration does not result in finite values (Camilli & Shepard, 1994). However, the RPU method accounts for some of these issues by implementing several variations on the calculation of difference between ICCs, some of which account for the empirical distribution of examinees. Additionally, probability differences are calculated at several points along the ability continuum such that integration is unnecessary. Further, likelihood ratio test procedures are time consuming and require specialized software. Likewise, DFIT can be complicated to implement especially since it requires the adequate calculation of covariance-variance matrices which current versions of the most popular calibration software does not currently provide.

Of all the model-based procedures, the plotting of IRT *b*-parameter estimates is the most easily performed method. Additionally, it directly compares to the delta plot method in terms of process as a statistical criteria can be adopted for flagging items that lie too far from the line of best fit in relation to other items' deviation. While several authors (e.g., Kolen & Brennan, 2004) suggest these bivariate plots and it is commonly used in practice, research to support its use is limited and no published studies could be found to compare its use to other methods or within simulated study conditions for IPD detection.

2.7 Impact on Ability Parameter Estimation & Classification

Of all the studies reviewed in Section 2.6 that compare aberrant item detection methods, delta plots are employed as one of the methods and findings are conclusive that all other methods have greater power for detecting IPD over administrations. However, several testing companies continue to use delta plots. Thus, the robustness of failing to identify and remove aberrant anchor items on the estimated scores and classification of examinees is important to determine and studies that begin this exploration are presented next.

Karkee and Choi (2005) found that the scaling equations used for the equating process changed once items were dropped from the set of anchor items based on model-based methods (i.e., in comparison to the delta plot method which flagged no items as aberrant), thereby changing the overall scale scores for students. Of the assessments examined, Algebra scores increased while Reading/Language arts scores decreased when aberrant anchor items were removed from the equating. Likewise, when Michaelides (2006) compared the performance of the delta plot method with the MH statistic for identifying aberrant anchor items across four assessments resulting in the MH statistic flagging more items; the effect of including or excluding the aberrant item(s) were as predicted. If the examinees in the Year 2 administration were performing higher than those in Year 1, including the item in the anchor led to a higher percentage of students being classified as proficient. Comparisons between the methods were made using a series of empirical data sets, and so it is impossible to ascertain which method was more accurate, only that the methods differed and lead to detectable differences in classification rates. It is also important to note that the predicted direction of classification error would also depend upon the direction of the drift for the detected items.

Kim and Nering (2007) conducted a simulation study that investigated the utility of the DFIT framework (i.e., the NCDIF index using item parameter replication method) in comparison to the delta plot method for evaluating equating items. Four conditions were explored; (1) one item with 0.10 *b*-shift, (2) one item with 0.50 *b*-shift, (3) two items with 0.10 *b*-shift, and (4) two items with 0.50 *b*-shift. Each of these four conditions were implemented both for a condition of equivalent groups and nonequivalent groups (i.e., 0.50 mean shift). A total of 10 replications of each were conducted with 5,000 simulated examinees. Findings indicated that the NCDIF index produced some modest gains (e.g., 3% classification improvement) over the delta analyses especially in the presence of nonequivalent groups.

Wells, Subkoviak, and Serlin (2002) studied the effect of drift on theta estimates via a simulation study where the number of test items (i.e., 40 & 80), number of examinees (i.e., 300 & 1,000), percent of items containing DIF (i.e., 5, 10, 15, & 20), and direction of DIF (i.e., uniform & non-uniform) were manipulated variables. Findings resulted in minimal effects on theta even when 20% of the items had drifted under normal distributional conditions. However, upon closer inspection, the direction of DIF did matter in terms of where the effects were found in the distribution. When the discrimination parameter drifted, the location of the examinees mattered such that those toward the extremes were affected more than those toward the mean. Further, when the difficulty parameter drifted, a uniform effect across the ability distribution was observed. The authors note that effects may be more severe in the context of shorter tests (e.g., such as anchor sets) and once drift has had the chance to compound over several years.

Hu, Rogers, and Vukmirovic (2008) used MS and SL scaling methods among others and under a number of conditions to find that including the aberrant items in the equating led to more systematic error in the equated scores, as would be expected. However, the effect of including/excluding the aberrant item on the classification of examinees was not explored.

Therefore, Sukin and Keller (2008, 2009b, 2010) and Keller, Egan, and Schneider (2010) studied the effect on examinee classification when the decision was made to remove or retain one or more aberrant anchor items. Several conditions were manipulated across these studies to include degree of aberrancy, ability distribution of examinees, number of equating items, number of items simulated as aberrant, scaling methods employed. The results indicated that the percent of correctly classified students was affected from a negligible to moderate degree depending on the study conditions. However, in all studies, the under- and over-classification (i.e., Type I & II error rates) of examinees was affected by item removal decisions such that when items were simulated as easier and items were flagged and removed from the anchor, over-classification errors increased. These affects were less noticeable when longer anchors were employed (e.g., 14 items) in the Keller, Egan, and Schneider (2010) study. Additionally, they found the RPU method to be the most successful in producing a purified anchor that resulted in the most accurate classifications when applied to empirical assessment data. The RPU method was also found to produce the fewest over-classifications among methods employed.

Finally, Keller and Wells (2009) used a 14-item internal anchor set within a 56-item assessment to explore differences in scaling constants and classification decisions as the percent of DIF items increased and as groups become nonequivalent. The authors report that negligible IPD (e.g., 2% classification errors) was observed when one item was simulated to have a -0.8 *b*-shift within the equivalent group condition. Otherwise, substantive IPD (e.g., 4-8% classification errors) was observed (i.e., one or more items simulated to have drift with the nonequivalent group condition or more than one item simulated to have drift with the equivalent group condition). These findings generalized across three differing sets of cut scores. As aberrant items were removed from the anchor, errors decreased. Additionally, these authors discovered that with a 0.06 or greater overall mean shift for the test difficulty significant classification differences existed. Overall, Keller and Wells (2009) recommended the removal of items with IPD so long as content representation of the anchor is not jeopardized.

While the magnitude of impact on estimated scores and classifications varies across studies reviewed, probably due to the specific study conditions employed, it is probable that given the potential compounding of IPD for items left undetected over test administrations will lead to increased error in classifications and thus a distorted assessment of the amount of true growth from one year to the next.

2.8 Conclusions Based on Review of Literature

The NEAT equating design is versatile in aiding the comparability of test scores from multiple test forms and occasions when groups taking these tests differ in their underlying ability distributions. Extensions can be made to implement the matrix-sampling design as well. However, research has shown that care must be put into the construction of the anchor set in order to appropriately scale test scores; and just as an assessment (i.e., operational) can contain items with DIF, so too can the anchor set. Consequences of using items in the anchor that contain year to year IPD, as a result of over exposure or familiarity with the testing conditions or differential opportunity to learn, can skew the resulting equating that takes place. However, in the absence of knowing the cause (i.e., DIF or bias), removal and inclusion decisions of aberrant items should be made with caution. Thus, it is imperative to assess which method and under what criteria for flagging have the most tolerable Type I error and power rates for detecting invariant items along with the impacts on classification for examinees as a result of practical decisions made about the inclusion or exclusion of such items.

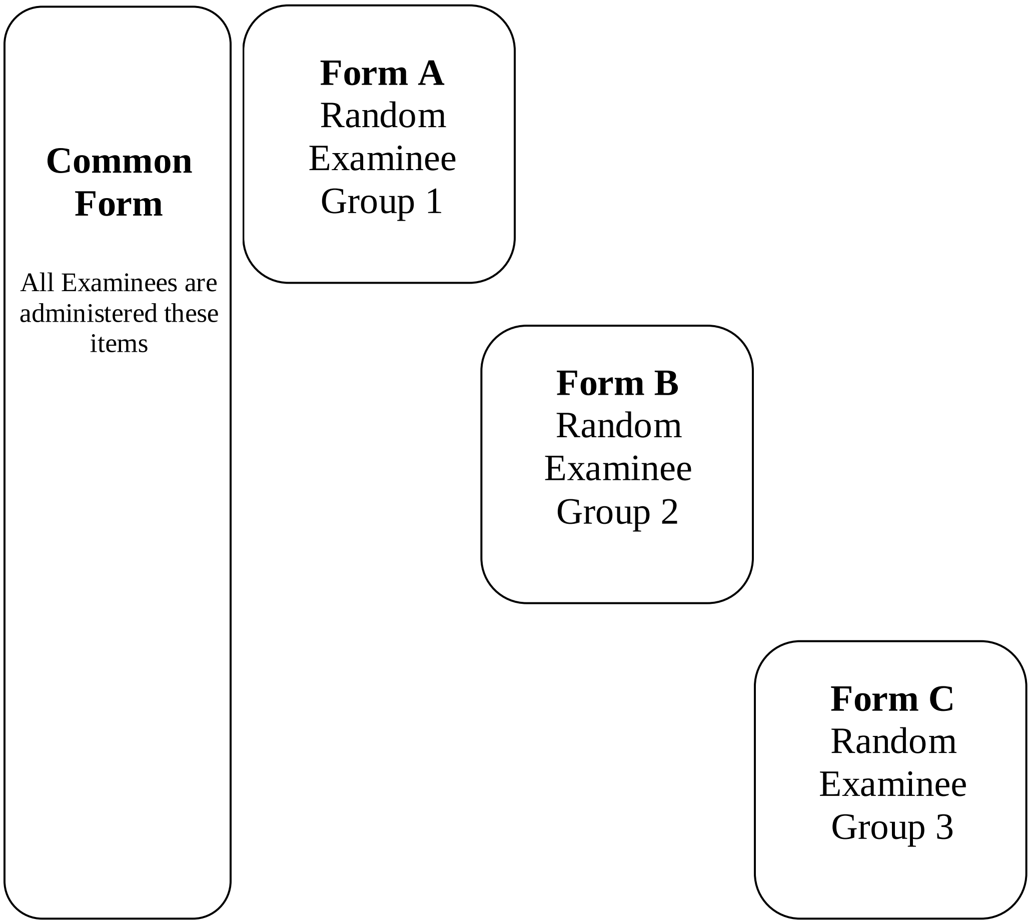


Figure 2.1. Matrix-Sampling Design Illustration

CHAPTER 3

METHODOLOGY

3.1 Overview

As the goal of this study is to provide additional and practical knowledge to the field of test scaling and equating such that test practitioners can increase the quality of their scaling and equating methods; this study will explore alternative methods for flagging aberrant (i.e., outlying) anchor items, both in the presence and absence of a distributional shift of examinees on ability. Additionally, how differences in the estimated parameters of outlying anchor items, whether the items are removed or retained for equating, interact with classification accuracy and ability estimation will be investigated. More specifically, the following questions will be addressed:

(1) Which methods for flagging outlying anchor items perform best to control Type I and II error rates?

(2) How does choosing a different method for flagging outlying anchor items impact the classification of examinees into performance groups?

(3) Does removing items from the anchor set used for equating of test forms due to IPD significantly mask the magnitude of growth when true growth is suspected?

Each of these questions will be answered via simulation data analyses informed by empirical item parameters. Following the simulation studies, item detection methods will be applied to empirical data and compared to the current practice followed for the statewide assessment data employed.

What follows is an outline of (1) a series of simulation studies proposed and (2) the application of recommendations to a statewide accountability assessment. The general procedural outline of the study follows:

1. Simulate item response data for two administrations of an exam containing both dichotomously and polytomously scored items and including zero, one, three, or five aberrant anchor items with varying initial item parameters;
2. Calibrate the items using separate parameter estimation with the three parameter logistic model (3-PLM) and graded response model (GRM) [PARSCALE: SSI, 2003];
3. Determine whether there are aberrant anchor items using the following methods: (1) delta plots, (2) plots of IRT *b*-parameter estimates, and (3) the RPU procedure;
4. Adjust evaluation criteria guidelines for methods if too few or too many items with IPD are flagged (i.e., conducted as a preliminary investigation);
5. Decide which aberrant item detection method in step 3 lead to the most accurately identified aberrant items;
6. Scale the two test forms using both sets of scaling constants derived (Stocking & Lord, 1983) with and without the flagged aberrant anchor item(s) [STUIRT: Kim & Kolen, 2004];
7. Apply appropriate scaling constants to the ability estimates obtained in the second administration;
8. Calculate root mean squared error (RMSE) between generated ability estimates and calibrated ability estimates for each of the conditions employed;

9. Classify the simulated examinees into performance categories based on the equated ability distributions obtained in step 6, both with and without the aberrant anchor item(s);
10. Compare the classification of the examinees in step 9, with the classification of the generating ability parameters for simulated examinees;
11. Decide which classification in step 9 lead to the most accurate classification;
12. Apply and compare methods with adjusted evaluation criteria from step 4 to archived statewide accountability data;
13. Report the efficacy of each of the detection methods and the potential implications of item removal and retainment decisions on examinee classifications (i.e., using a difference that matters-DTM-criterion); and
14. Determine whether growth is masked in the condition where it is simulated by comparing classification differences between conditions where aberrant items are removed and classifications assigned based on the condition of no drift.

The rest of this chapter, details the IRT models used, the simulation study conditions and procedures as outlined above, and a description of the operational assessment data and procedures used.

3.2 IRT Models

Three IRT models are employed in this study; (1) 3-PLM, (2) GRM, and (3) multidimensional. The first two assume unidimensionality and are used for item calibration, ability estimation, and ability parameter generation for the second assessment's administration. The multidimensional IRT model is used for generating the initial ability and opportunity to learn (OTL) parameters for simulated examinees during the first administration of the assessment. Each of these models and their utility within this study are described next.

3.2.1 3-PLM

The 3-PLM is used to calibrate and provide ability estimates for examinees for multiple-choice (MC) and binary scored short answer (SA) items. With the SA items, the *c*-parameter will be fixed to zero, as correctly responding to an item of this type by chance alone is often unlikely. The formula for the model appears in equation 1,

$$P_i(\theta_j) = c_i + (1 - c_i)\frac{\exp(Da_i(\theta_j - b_i))}{1 + \exp(Da_i(\theta_j - b_i))} \tag{1}$$

where *i* indexes the items, *j* indexes examinees, *a* represents the item discrimination parameter, *b* represents the item difficulty parameter, *c* is the pseudo-guessing parameter, and *D* is a normalizing constant equal to approximately 1.7.

3.2.2 GRM

The GRM (Samejima, 1969) is used to calibrate and provide ability estimates for examinees for constructed response (CR) items that are scored using more than two points along a scale (i.e., polytomously scored). Thus, an item is scored in *k*+1 graded categories, where each *k* has its own threshold and can be treated as a binary item with each adjacent category. Therefore, a polytomous item with *k*+1 categories can be modeled as a set of item category threshold curves (ICTC) using *k* 2-PLM formulations as in equation 2:

$$P^*_{ik}(\theta_j) = \frac{\exp(Da_i(\theta_j - b_i + d_{ik}))}{1 + \exp(Da_i(\theta_j - b_i + d_{ik}))}, \tag{2}$$

where *i* indexes the items, *j* indexes examinees, *k* indexes thresholds, *a* represents the item discrimination parameter, *b* represents the item difficulty parameter, *d* represents a category step parameter, and *D* is a normalizing constant equal to approximately 1.7. After computing *k* ICTCs, *k*+1 item category characteristic curves (ICCC) are then calculated by subtracting adjacent ICTC curves:

$$P_{ik}(\theta) = P^*_{i(k-1)}\big(\theta_j\big) - P^*_{ik}(\theta_j), \qquad (3)$$

where P_{ik} represents the probability that the score on item *i* falls in category *k*, P^*_{ik} represents the probability that the score on item *i* falls above the threshold *k*. Additionally, it should be noted that $P^*_{i0} = 1$ and $P^*_{i(k+1)} = 0$.

3.2.3 Multidimensional IRT

The multidimensional IRT model is used to generate initial ability parameters, θ and η, for the first year's test administration. The θ parameter is meant to represent the simulated examinee's ability on the construct of interest (i.e., the one intended to be measured); while the η parameter represents the simulated examinee's magnitude or degree of opportunity to learn (OTL) the content presented on a subset of the assessment simulated to be neglected from instructional focus. Equation 4 describes the multidimensional IRT model used (de Ayala, 2009):

$$P_i(\theta_j) = c_i + (1 - c_i)\frac{\exp(\underline{a}_i^{'}\underline{\theta}_j + \gamma_i)}{1 + \exp(\underline{a}_i^{'}\underline{\theta}_j + \gamma_i)}, \qquad (4)$$

where *i* indexes the items, *j* indexes examinees, $\underline{\theta}$ represents the vector of ability parameters (i.e., θ and η in this case), $\underline{a'}$ represents the transposed vector of item discrimination parameters, and γ represents -($a_{i1}b_{i1}$+ $a_{i2}b_{i2}$) for the specific case where the θ vector contains two elements and *a* represents the discrimination parameter and *b* represents the difficulty parameter of θ and η, respectively. In the instance where polytomous items are modeled using both θ and η parameters, the GRM can be adapted in a similar fashion. However, for the sake of parsimony, polytomously scored items are not modeled as containing any drift. Thus, only the 3-PLM extension will be used for MC items and the 2-PLM (i.e., *c* is fixed to zero in equation 4) for binary scored SA items. For items where OTL is not suspected to be acting differentially between years, the *a*-parameter for η will be set to zero.

3.3 Simulation Study

The following sections provide a detailed description of the methods to be invoked for the simulation study conducted as outlined in Section 3.1.

3.3.1 Test Design

Four sets of 43-item (33 MC, 3 two-point SA, 4 three-point CR, and 3 four-point CR) administrations of parallel assessments are simulated using the NEAT design. A 40-item (36 MC and 4 four-point CR) external anchor is used within a matrix-sampled design such that items are spread across a total of four paired forms. Thus, each simulated examinee sees nine MC and one CR anchor item in addition to the 43 operational items. This design is illustrated in Figure 3.1 and is meant to approximate the design of the science assessment of the operational statewide testing program used for the second part of this study.

3.3.2 Parameters

The item parameters used for generating the item responses were obtained from an operational statewide testing program and can be found in Table A.2 for the first administration and Table A.3 for the second administration located in Appendix A. A total of 126 items (102 MC, 6 SA, 8 three-point CR and 10 four-point CR) were simulated. A multidimensional three parameter IRT model, 3-PLM, and GRM are used to simulate the item responses as described in Section 3.2 and as pictured in Figure 3.1. Eight thousand examinees were simulated for each of the four administrations using *SPL2K* (2010) software. In the first case, examinees were drawn from two conditions of a bivariate normal distribution with a mean vector equal to zero and a covariance matrix equal to $\begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}$ or equal to $\begin{bmatrix} 1 & 0.6 \\ 0.6 & 1 \end{bmatrix}$. Thus, multidimensionality is incorporated into the overall theta values of the first administration to represent the construct of interest (θ) and opportunity to learn (OTL; η), or the lack thereof, as a nuisance parameter. While there is no real research to indicate how highly correlated the construct and indicators of OTL are for science assessments, it may be reasonable to expect an estimate of around 0.8 or 0.6. Therefore, these values will be used as examples to begin research in this area.

In the second case, examinees were modeled with unidimensional ability estimates. Examinee ability parameters were drawn from a N(0,1) distribution for the first condition to simulate equivalent groups between years. For the second condition, growth was simulated such that examinee ability parameters were sampled from a negatively skewed distribution with a mean of approximately 0.2 and a standard deviation of around 0.8. It is typical that as the distribution becomes more negatively skewed, the standard deviation also decreases. Thus, in the second condition, both the mean and standard deviations have shifted to reflect the type of distributional shift expected in the presence of year-to-year growth on the construct of interest. Additionally, as growth is more likely to be found in the lower portion of the distribution, examinees simulated to have thetas less than -0.50 will be shifted upwards by 0.50 and simulated examinees with thetas greater than -0.5 will be shifted upwards by 0.15 for their respective counterparts; thus, producing skewness and kurtosis magnitudes of 0.17 and 0.50, respectively. Descriptive statistics for all simulated examinee populations can be found in Appendix A (Table A.4).

3.3.3 Aberrant Items

Five conditions of aberrancy are explored such that for any given study condition, between zero and three items were chosen and simulated to be aberrant. These items have parameters spread across the item parameter continuum and mimic the properties often found on operational tests. The items were made aberrant by shifting both the *b*- and *a*-parameters (*ab*-drift) between administrations using four combinations of shift magnitudes. The *b*-parameters were shifted by two different values, -0.5 and -0.8, to simulate two degrees of aberrancy. These values were chosen to reflect the threshold with which a difference is likely to occur (Han & Wells, 2007) and a moderate DIF condition as suggested by Jodoin and Gierl (2001), respectively. Likewise the *a*-parameters were shifted by values of -0.3 and -0.7, as suggested by C. S. Wells (personal communication, February 4th, 2010) and found to be within the range of *a*-parameter shifts observed in previous studies of “mixed” DIF (e.g., Narayanon & Swaminathan, 1996; Rogers & Swaminathan, 1993). Each of these magnitudes is crossed to produce four combinations of IPD. Table A.3 and Figures A.1 through A.5 in Appendix A provide all resulting parameter estimates and ICCs based on these adjustments for linking items manipulated in this way.

Further, placement of these items are strategically chosen to explore the power of each of detection methods to flag aberrant items dependent upon their placement along the ability continuum. The following five schemes are implemented:

1. *Null Condition*: No items are simulated to be aberrant to explore Type I error rates in the presence of a pure growth effect.
2. *1 Hard Item*: An item that straddles the highest cut score (i.e., in the first administration, the item would indicate knowledge at a proficient level and in the second administration, the item would indicate knowledge at an advanced, distinguished, or above proficient level – provided the parameter estimates were accurate) is simulated to be aberrant.
3. *3 Spread Items:* Three items are simulated to be aberrant such that one item straddles each of the cut scores.
4. *3 Moderate Items*: Three items are simulated to be aberrant such that two straddle the second cut score and one straddles the highest cut score.
5. *5 Spread Items*: Five items are simulated to be aberrant such that two items straddle the lowest cut score, two items straddle the second cut score, and one item straddles the highest cut score.

3.3.4 Aberrant Item Detection

The following section details the technical procedures for each of the aberrant item detection methods employed in this study to include the; (1) delta plot, (2) IRT *b*-parameter plots, and (3) the RPU procedure. Of all the available methods, the chosen methods were decided upon based on the study conditions and the review of the literature. Detection methods were eliminated for study based on consistent findings of lack of power to detect non-uniform DIF (e.g., MH), previous studies showing lack of item flagging consistency across replications (i.e., TCC Inverse), inflated Type I error rates (i.e., Lord’s chi-square), or lack of practicality operationally (e.g., DFIT and likelihood ratio test). The delta plot method is included as a control condition since a review of several testing programs revealed that many statewide testing programs use this method for analyzing their anchor items used for equating.

3.3.4.1 Delta Plot

The delta plot method (Huff & Hambleton, 2001; Kolen & Brennan, 2006), sometimes referred to as the transformed item difficulty (TID) index (Angoff & Ford, 1973; Camilli & Shepard, 1994) relies upon the item difficulty values (i.e., *p*-values) for the common items between the referent (i.e., examinees taking Administration 1) and focal (i.e., examinees taking Administration 2) groups. These *p*-values are first transformed onto the inverse normal function and then further transformed into delt*a*-values (Holland & Thayer, 1985) using equation 5. From equation 5, one minus the *p*-value is converted to a *z*-score via a *p*-to-*z* transformation using the inverse of the normal cumulative function (Dorans & Holland, 1993). This transformation is conducted to remove the curvilinear relationship between each of the subgroups’ *p*-values produced by floor and ceiling effects (i.e., *c* to 1) where the smallest differences would be seen for the hardest and easiest items. Finally, the *z*-scores are re-scaled using the delta scale score linear transformation. Delta values are then plotted against each other after separate calibration has taken place. Lastly, a best-fitting line (i.e., one that minimizes the perpendicular deviation such that a symmetric solution is produced) is drawn through these values and an anchor item is flagged for investigation if the item lays more than two or three standard deviations from this line. While three standard deviations are most often used as the critical value in practice, it has often been found that this criterion results in low power (Keller, Egan, & Schneider, 2010; Sukin & Keller, 2008; Sukin & Keller, 2009b; Karkee & Choi, 2005). Thus, this study will employ a more liberal criterion of two standard deviations. Additionally, a trial study was performed for this study using three standard deviations as the criterion which resulted in confirmation of earlier findings. These results can be found in Table A.5 of Appendix A.

$$\Delta = 13 + 4z_{(1-p)} \tag{5}$$

Further, the delta plot method can be easily extended to the identification of polytomous items that experience drift by rescaling the item responses to a scale of zero to one and plotting the plots along with the dichotomous items.

3.3.4.2 IRT *b*-Parameter Plots

Plots of IRT *b*-parameter estimates (Huff & Hambleton, 2001) are conducted in a very similar fashion to delta plots; such that item difficulty estimates (i.e., *b*-values) for the common items between the referent (i.e., examinees taking Administration 1) and focal (i.e., examinees taking Administration 2) groups are obtained and theoretically plotted against each other after separate calibration has taken place. A best-fitting line (i.e., one that minimizes the perpendicular deviation such that a symmetric solution is produced) is drawn through these values and an anchor item is flagged for investigation if the item lays more than two or three standard deviations from this line. Based on a preliminary study for this research two standard deviations were chosen due both to the lack of power and extremely modest Type I error rate when using three standard deviations as the criterion. Results for this trial study can be found in Table A.6 in

Appendix A. Further, this method is easily extended to the identification of polytomous items that experience IPD in one of two ways; (1) using the global *b*-parameter estimate for comparison or (2) using the threshold estimates for each score point and deciding upon a decision rule of when to exclude or include polytomous items for equating when one or more score points are flagged for IPD. For this study, the first method is implemented as polytomous items are not simulated to function differentially and no well-known and researched decision rules for implementing the second method exist.

3.3.4.3 RPU Method

The RPU method is unique as it employs seven statistics based on the examination of the difference between item characteristic curves (ICCs) between assessment administrations. Statistics invoked include; (1) average signed difference in estimated probability, (2) average unsigned difference in estimated probability, (3) root mean squared difference, (4) weighted average signed difference in estimated probability, (5) weighted average unsigned difference in estimated probability, (6) weighted root mean squared difference, and (7) maximum signed difference. Each of these statistics is presented in equations 6.1 through 6.7, respectively:

$$\frac{\sum_{j=1}^{51}(P_{i2}(\theta_j)-P_{i1}(\theta_j))}{51}, \tag{6.1}$$

$$\frac{\sum_{j=1}^{51}\left|P_{i2}(\theta_j)-P_{i1}(\theta_j)\right|}{51}, \tag{6.2}$$

$$\sqrt{\frac{\sum_{j=1}^{51}\left(P_{i2}(\theta_j)-P_{i1}(\theta_j)\right)^2}{51}}, \tag{6.3}$$

$$\frac{\sum_{j=1}^{51} n_j(P_{i2}(\theta_j) - P_{i1}(\theta_j))}{51 * n}, \tag{6.4}$$

$$\frac{\sum_{j=1}^{51} \left| n_j(P_{i2}(\theta_j) - P_{i1}(\theta_j)) \right|}{51 * n}, \tag{6.5}$$

$$\sqrt{\frac{\sum_{j=1}^{51} n_j\left(P_{i2}(\theta_j) - P_{i1}(\theta_j)\right)^2}{51 * n}}, \tag{6.6}$$

and

$$\max_j \mid P_{i2}(\theta_j) - P_{i1}(\theta_j) \mid; \tag{6.7}$$

where $P_{ik}(\theta_j)$ is the probability of a correct response for item *i* in year *k* for a theta value of θ_j, where theta is evaluated between -2.5 and 2.5 in increments of 0.1. This results in 51 points along the theta continuum where the ICC differences are calculated. Additionally, n_j represents the number of examinees in the interval *j*, and *n* is the total number of examinees for equations 6.3 through 6.6 where weighted differences are calculated. Typically, differences between 0.07 and 0.10 are considered to be moderate (Keller, Egan, & Schneider, 2010) for the first six statistics. Values that are higher than 0.10 are considered to be large. Likewise, for the maximum signed difference (equation 6.7), moderate differences are declared when values range between 0.125 and 0.15 with differences greater than 0.15 considered as large. Keller, Egan, and Schneider (2010) explain that when used operationally, an item becomes a candidate for removal when it is identified as having large differences for four of the seven statistics considered. However, in employing the more conservative criterion for a trial analysis, the detection power was observed to be low. These results can be found in Table A.7 of Appendix A. Thus, this study will employ the moderate effect sizes of 0.07 for equations 6.1 through 6.6 and 0.125 for equation 6.7. An item will be removed from the anchor when four of the seven statistics surpass these thresholds. To adapt this method to polytomous items, individual thresholds will be examined in similar ways and items will be flagged for aberrancy so long as at least one threshold surpasses four of the seven effect size magnitudes.

3.3.5 Scaling

Stocking and Lord's (1983) test characteristic curve method for scaling will be implemented using STUIRT (Kim & Kolen, 2004) software as past research has shown that results differ very little between methods employed on the final classifications of examinees (e.g., Sukin & Keller, 2008, 2009b). Additionally, in practice, this method is common place. Using this method, *A* and *B* scaling constants are obtained and applied (Kolen & Brennan, 2004) such that the new parameter values for Administration 2's assessment placed onto the scale of Administration 1 become a^* (equation 7), b^*(equation 8), c^*(equation 9), and θ^* (equation 10):

$$a^* = \frac{a}{A} \tag{9}$$

$$b^* = Ab + B \tag{10}$$

$$c^* = c \tag{11}$$

$$\theta^* = A\theta + B \tag{12}$$

3.3.6 Summary of Conditions

The following outlines the simulation study conditions:

- 2 Administration 1 Ability Distributions (bivariate normal with mean of zero and covariance matrices of $\begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}$ and $\begin{bmatrix} 1 & 0.6 \\ 0.6 & 1 \end{bmatrix}$, respectively);
- 2 Administration 2 Ability Distributions (N(0,1) and negatively skewed growth distribution of (0.26, 0.89, 0.17, 0.50));
- 4 IPD Magnitude Shifts (2 *b*-parameter shifts (-0.5 & -0.8) crossed with 2 *a*-parameter shifts (-0.3 & -0.7);
- 5 Aberrant Item Schemes (Null Condition, 1 Hard Item, 3 Spread Items, 3 Moderate Items, & 5 Spread Items); and
- 4 IPD Detection Conditions (None, Delta Plots, IRT *b*-parameter plots, & RPU Method).

Partially crossed (i.e., accounting for the fact that when no items are simulated as aberrant, the parameter shifts can not occur), this produces a 2 x 2 x 4 x 4 design for 64 study conditions and an additional 4 control conditions resulting in a total of 68 study conditions that will be assessed using each of the conditions for IPD detection. A total of one hundred replications will be performed for each condition.

3.4 Evaluation Criteria

The following section outlines the evaluation criteria for the simulation study. First, three different item detection methods were investigated for accuracy of detecting aberrant items. Second, the recovery of ability estimates for each of the groups is assessed. Next, the accurate classification of examinees was explored for both including the aberrant items and excluding them in the equating. Finally, based on the condition for which growth is simulated, expected growth rates are calculated and assessed in comparison to the condition where no drift is simulated.

3.4.1 Aberrancy Detection

This section outlines the criteria used for each of the flagging methods as summarized in Table 3.1 along with an examination of Type I error rates and power.

3.4.1.1 Type I Error

The proportion of items that are incorrectly flagged as expressing IPD are averaged and reported for each of the methods across replications. It is expected that Type I error rates are maintained at 0.05 or less. Type I error rates are reported and interpreted first, as any inflation in Type I error rates are expected to artificially increase the power.

3.4.1.2 Power

To determine which IPD detection method most accurately detected items simulated under varying conditions of type and magnitude of aberrancy, the proportion of times each of the items were classified as aberrant were calculated for each of the conditions implemented.

3.4.2 Ability Distribution Recovery

The distributions of each group are assessed in terms of how well the expected ability parameters are recovered based upon the item removal and retainment decisions and in comparison to the ability parameters generated. These distributions will be assessed graphically by plotting the average discrepancy between the generated thetas (*ø*) and the study condition (*s*) for each 0.1 interval on the theta scale where *j* represents the simulated examinee and *J* represents the total number of simulated examinees (i.e., 32,000). These differences will be summarized using the root mean squared error (RMSE) statistic, as presented in equation 13, averaged over replications (r):

$$RMSE_r = \sqrt{\frac{\sum_{j=1}^{J}(\theta_{\emptyset} - \theta_s)^2}{J}}\,. \tag{13}$$

3.4.3 Examinee Classification

This section outlines how simulated examinees are classified into proficiency categories.

3.4.3.1 Cut Scores

To simulate a common practice in many testing programs, examinees are classified into one of four performance categories, based on three cut scores. The cut scores were chosen to accommodate the aberrant item schemes such that items would straddle cut scores in the appropriate manner based on item parameters obtained in practice. Cut scores of -0.75, 0.00, and 1.50 on the theta metric were chosen to classify examinees.

3.4.3.2 Classification

To compare the effect of aberrancy on the different scaling methods, classification accuracy was determined for each of the methods both with and without the aberrant anchor items. Therefore, for each examinee, three classifications were determined:

1. True Classification—the classification of the examinee based on the generated theta parameter.
2. Aberrant Classification—the classification of the examinee obtained when the aberrant items are left in the anchor for equating.

3. Purified Classification—the classification of the examinee obtained when the flagged aberrant items are removed from the anchor for equating.

Using these three classifications, four contingency tables are created for each method and each replication. Comparisons are made between the true classification and the aberrant classification (i.e., one table) as well as the true classification and the purified classification (i.e., three tables, one for each IPD detection method). Examinees are placed into one of sixteen categories as shown in the example presented in Table 3.2. Additionally, the black categories are collapsed and indicate a correct classification of examinees, the gray categories are collapsed and represent an over-classification (i.e., Type II error) of examinees, and the white boxes are collapsed and represent an under-classification (i.e., Type I error) of examinees.

Thus, for each method, the percent of accurately classified, over-classified (i.e, false-positives), and under-classified (i.e., false-negatives) examinees are computed for each replication and averaged over replications. The effects that aberrant anchor items had on the accuracy of classification of examinees into performance categories are thereby determined.

Comparisons are made using a difference that matters (DTM) effect size, such that less than a 1.8% difference in classifications between detection method outcomes is considered small; as Keller and Wells (2009) defined a conservative estimate of a DTM as differences larger than those expected due to rounding to the first decimal place of the theta value. Therefore, any differences larger than 1.8% are considered to be practically significant.

3.4.4 Growth Expectations

This section outlines how growth is assessed for determining whether the act of removing items that are flagged as aberrant mask growth when opportunity to learn is suspected as the primary reason for IPD. Growth is calculated by assessing the percentage of examinees that obtain a higher classification (i.e., one point is awarded to each examinee pair that exhibits an increased proficiency level) when assessed within the negatively skewed distribution in comparison to their simulated comparison examinees within the normally distributed population for which growth is not simulated. Thus, the condition for which no drift is simulated is used as the baseline for making comparisons. Again, the absolute valued DTM threshold of 1.8% is used to indicate practically significant differences. In cases where this difference threshold is exceeded, growth is considered to be over estimated if the difference is positive and under estimated if the difference is negative. In all other cases, the growth is considered to be appropriately estimated.

3.5 Empirical Application

This section describes empirical data from a statewide assessment used for accountability purposes. This study looks at the effects of anchor item removal and retainment decision differences between item parameter drift detection methodologies and as applied to data where a distributional shift is expected. These distributional shifts were expected to occur in the science assessment due to changes in curricular emphasis as indicated by an evaluation of previous year's assessment results. A more detailed description of the hypothesis, test design, response data, and methods follows.

3.5.1 Hypothesis

As noted previously, the science test in this assessment system provides a unique opportunity to examine the effects of opportunity to learn (OTL). It was recognized that a strand of the science curriculum was not being emphasized in the instruction. As such, examinees were not performing well on this strand. Once this trend was detected, the strand was then emphasized in instruction, which would likely lead to changes in the performance of examinees on items within this strand. These items are likely to exhibit IPD. Additionally, it is hypothesized that the dimensionality of the test would likely change between administrations as well such that any multidimensionality detected in the first assessment would become more weak in the second administration as the implementation of instruction becomes more aligned with curriculum and testing frameworks.

3.5.2 Test Design & Data Description

The statewide science assessment is given in grades 4, 8, and 11 and is composed of three sections such that the first section consists of 25 multiple-choice items and three constructed response items; the second section consists of 26 multiple-choice items and three constructed response items; and the final section consists of a combination of 2-point short answer and 3-point constructed response items. Thus, there are total of 64 items. Additionally, there are a total of four forms and the assessment schematic described includes both field-test and matrix-sampled items. Only the common items across the four forms are used for scoring examinees. For the sake of this study, the field test items are ignored and thus removed from the data analysis. Lastly, the matrix-sampled items, of which there are nine MC items and one 4-point CR item per test form, are used for this analysis. The assessment design is displayed in Figure 3.1. Table 3.3 provides the total number of examinees tested by grade. Table 3.4 provides descriptive statistics for the ability parameters for each of the assessments investigated in this study.

3.5.3 Methodology

The basic methodology of the simulation study will be followed for the empirical data. Since the delta plot method using three standard deviations as a criterion is used operationally for identifying items that exhibit IPD, this condition will serve as the comparison for evaluating the classification of examinees. Therefore, the analysis will consist of the following steps. First, each of the three methods: delta plot (i.e., using both a criterion of 2 and 3 standard deviations), IRT *b*-parameter plots, and the RPU method as described previously in the simulation study, will be applied to the science assessment data for flagging items possessing IPD. Secondly, the flagged items will be removed from the anchor set and equating will be performed without them. Third, examinees will be classified into proficiency categories based on this new equating. Fourth, differences between original classifications (i.e., based on the delta plot analysis using three standard deviations as the criterion) and subsequent classifications (i.e., based on the equating results using the delta plot analysis using two standard deviations as the criterion, IRT *b*-parameter plots, and the RPU method as detection methods) will be explored using the DTM criteria described in Section 3.4.1.2. Lastly, any discrepancies between methods and their implications for practice will be discussed.

Table 3.1. Aberrant Anchor Item Flagging Criterion for Different Methods

Methods	Flagging Criteria	Source
Delta Plots	Perpendicular distance > 2 SD* from the line of best fit	Angoff & Ford (1973)
IRT *b*-parameter plots	Perpendicular distance > 2 SD* from the line of best fit	Kolen, & Brennan (2004); Huff & Hambleton (2001)
RPU Method	Items with four of seven statistics that exceed thresholds of 0.07 for equations 6.1 through 6.6 and 0.125 for equation 6.7	Keller, Egan, & Schneider (2010)

*SD: Standard deviation

Table 3.2. Contingency Table for Four Performance Levels, Aberrant Classification vs. True Classification

		Aberrant Classification			
		1	2	3	4
True Classification	1	■	▨	▨	▨
	2		■	▨	▨
	3			■	▨
	4				■

Table 3.3. Number of Examinees Tested by Grade and Administration

Grade	Number of Examinees Administration 1	Number of Examinees Administration 2
4	32,227	30,500
8	34,823	33,750
11	32,260	32,647

Table 3.4. Descriptive Statistics of Theta Parameters, Operational Assessment

Grade	Administration	Mean	Median	Standard Deviation	Skewness	Kurtosis
4	1	-0.076	-0.034	0.958	-0.267	0.016
	2	-0.026	0.037	0.939	-0.408	0.092
8	1	-0.054	-0.002	0.973	-0.288	0.074
	2	-0.085	-0.012	0.983	-0.300	-0.154
11	1	-0.024	0.056	0.967	-0.358	0.005
	2	0.061	0.145	0.926	-0.343	-0.211

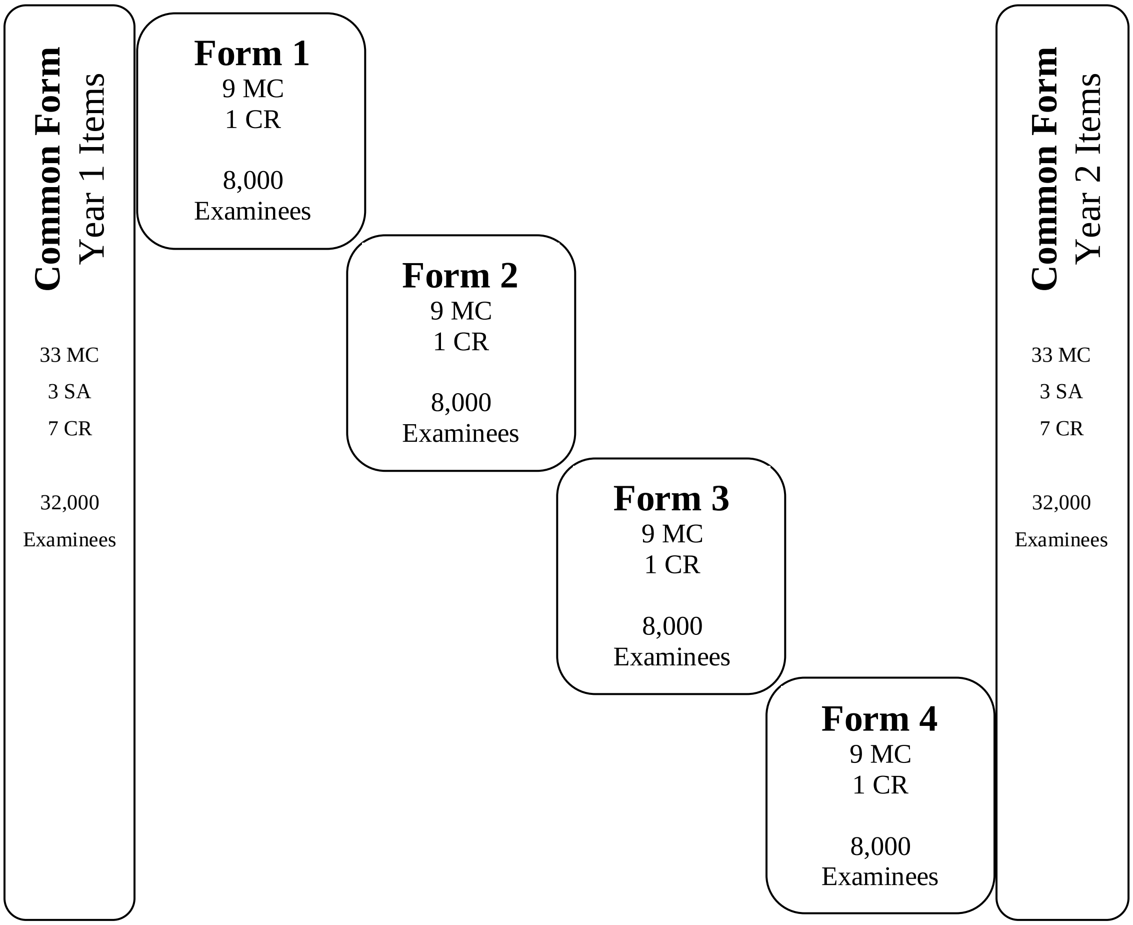


Figure 3.1. Matrix-Sampling Design Employed for Simulation Study

CHAPTER 4

RESULTS

4.1 Overview

This study consisted of both a simulation and an empirical data analysis. Results for each will be discussed in turn. The simulation study included five variables (i.e., administration 1 theta correlation condition, aberrant item scheme, degree of aberrancy combination, ability distribution shift condition, and aberrant item detection method) resulting in a total of 260 study conditions. Each study condition was replicated 100 times. For each of the replications, four sets of scaling constants were obtained based on each of the aberrant item detection methods employed. Calibrations for items occurred separately for each administration and all calibrations resulted in converged solutions; thus allowing all replications to be used within the final analysis. Results for the simulation study are divided into four sections: (1) aberrancy detection, (2) ability parameter recovery, (3) examinee classification, and (4) growth expectations and are summarized in the tables and figures as described next. Further, it is important to note that all reporting of results are averaged across the 100 replications employed for each set of conditions.

While a summary of the placement of tables and figures within this report appear here, they will also be referenced later when appropriate, throughout this report. The results for aberrancy detection, which includes Type I error and power rates, are displayed in Tables 4.1 through 4.6. A table appears for each of the detection methods and each of the administration 1 conditions. Power is further summarized in Figures 4.1 through 4.12 and in Appendix B by averaging the detection power across items simulated as aberrant. The RMSE statistic is used to describe the ability parameter recovery rate and results can be found in Table 4.7 for all study conditions employed. Additionally, Appendix C presents the average discrepancy for 51 intervals between -2.5 and 2.5 on the theta scale. Examinee classification is summarized in Tables 4.8 through 4.10 for which each table represents accurate, under-, and over- classifications, respectively. More detailed classification contingency tables appear in Appendix D. Figures 4.13 through 4.28 display accurate classification rates for each of the detection methods using the DTM statistic in relation to the absence of a detection method. Further detail is provided in the figures within Appendix E resulting in a set of plots for accurate, under-, and over-classification rates using the DTM statistic in relation to the condition for which no IPD was simulated for items. Finally, Table 4.11 presents the assessment of growth in comparison to a condition for which no drift is simulated.

The empirical data analysis applies the IPD detection methods employed in the simulation study to archived science achievement data for three grade levels. Over 30,000 examinees are available for each of the grade levels and a matrix-sampling equating design is employed. Results are divided into three sections: (1) item parameter drift analysis, (2) equating summary, and (3) proficiency classification and are summarized in the tables and figures presented next. These tables and figures will also be referenced throughout this report when appropriate. IPD results are presented in Tables 4.12 through 4.20. Scaling constants obtained resulting from the item inclusion and exclusion decisions based on the various IPD detection methods appear in Table 4.21. Resulting classifications into proficiency categories are summarized in Tables 4.22 though 4.24.

Results for the simulation study are presented first, followed by the empirical data analysis.

4.2 Simulation Study

4.2.1 Aberrancy Detection

Depending on the condition employed, up to five items were simulated to have varying degrees of aberrancy. In the assessment of power, multiple interaction effects were observed. Of these items, the one with the largest initial *b*-parameter (i.e., the most difficult item) was detected the most frequently by all methods and under most degrees of aberrancy. Further, under most conditions, the Type I error rate did not exceed 0.05. The exception occurred when ability parameters were simulated to be negatively skewed and the RPU detection method was employed. More detailed results for power and Type I error will be discussed next.

4.2.1.1 Power

Most often, the average power across all methods of item detection was only sufficient (i.e., exceeding 0.85) for the condition for which only one difficult item was simulated as aberrant. This statement fails to be true for conditions as listed; (1) RPU method under all conditions that exhibit a -0.3 *a*-shift, (2) RPU method with a -0.7 *a*-shift under all conditions for which the second administration exhibits a negatively skewed distribution, (3) *b*-parameter plot method for the lowest degree of aberrancy (i.e., -0.3 *a*-shift and -0.5 *b*-shift) when no distributional shift is employed and administration 1 thetas are correlated 0.6, and (4) *b*-parameter plot method for a combination of -0.7 *a*-shift and -0.5 *b*-shift aberrancy for which there is a negatively skewed distribution for the second administration and thetas are correlated 0.6 for administration 1. The figures presented in

Appendix B display the average detection rate (i.e., Type I eror and power) across all items within each aberrant item scheme and for each detection method. The Type I error plots are for items that were not simulated with IPD and the power plots are for items that were simulated with IPD. A set of plots appears for each administration 1 condition (i.e., 0.6 correlated thetas or 0.8 correlated thetas) by IPD combination condition (i.e., shifts of -0.3 and -0.7 for the *a*-parameter fully crossed with shifts of -0.5 and -0.8 for the *b*-parameter). Two other instances for which power is sufficient occurs when the delta plot method is employed, there is no shift in the ability distribution between administrations, and the degree of aberrancy is such that there is a -0.3 *a*-shift and a -0.8 *b*-shift as displayed in Figure 4.2.

In addition to the overall power inadequacies, additional trends are difficult to describe as multiple interactions among conditions and detection methods employed exist as can be observed in Figures 4.1 through 4.12. The most conclusive finding is that for all conditions for which one item is simulated as aberrant, the delta plot method detects it with perfect power. Otherwise, there are too many interdependencies among conditions to make any conclusive statements regarding these findings. It may be appropriate to claim that as the administration 1 and 2 distributions diverge, the rate at which items are detected for aberrancy increases. However clear exceptions would include the following; (1) when the RPU method is employed, only one item is simulated as aberrant, and the *a*-shift is simulated as -0.7, (2) when the *b*-parameter plot method is employed, 3 items are simulated as aberrant such that the items are spread throughout the *b*-parameter continuum, and the drift is described by an *a*-shift of -0.7 and a *b*-shift of -0.5, and (3) when the delta plot method is employed, three items are simulated as aberrant such that the items are spread throughout the *b*-parameter continuum, and the drift is described by an *a*-shift of -0.3 and a *b*-shift of -0.8.

Trends within detection methods and for individual items can be more readily described using Tables 4.1 through 4.6.

4.2.1.1.1 Delta Plot

Regardless of condition, the manipulated anchor item that was the most difficult was always detected with perfect power. This is not true for any other items. However, interactions do exist among the conditions and items. Such interactions will only be discussed for items for which power is considered to be sufficient. First, it is noteworthy that regardless of the correlation between administration 1 and 2 thetas, the same items are flagged at similar rates across all study conditions. Further, when no distribution shift is present, the only other item that is sufficiently flagged is the moderately hard item for the condition where three items are spread across the difficulty continuum and for which there is a -0.3 *a*-shift and a -0.8 *b*-shift. When a distribution shift is present, items sufficiently flagged include the following; (1) one of the moderately hard items for the condition where three items of moderate to hard difficultly are simulated as aberrant and for which there is a -0.3 *a*-shift and a -0.8 *b*-shift, (2) an easier item for the condition where five items of varying difficulty are simulated as aberrant and for which there is at least a -0.7 *a*-shift, and (3) one of the moderately hard items for the condition where five items of varying difficulty are simulated as aberrant and for which the *a*-shift is equal to -0.3 and the *b*-shift is equal to -0.8. Overall, the delta plot method's power for detecting IPD was not affected by the degree of correlation between the first administration's ability parameters, but increased detection power was observed when a shift in the distributions was simulated. The amount of IPD and number of items simulated as aberrant showed inconsistent patterns.

4.2.1.1.2 IRT *b*-Parameter Plots

The manipulated anchor item that was the most difficult was the most consistently detected item across conditions. However, multiple interactions were observed where the detection rate for this item was not sufficient as displayed in Tables 4.3 and 4.4. For both conditions of correlated thetas for administration 1, the only other item sufficiently detected was an easier item within the condition for which five items were simulated as aberrant and the largest amount of drift was simulated (i.e., -0.7 *a*-shift and -0.8 *b*-shift). Under all other conditions, power never exceeded 0.85. This method did not seem to be affected consistently by any of the study conditions employed.

4.2.1.1.3 RPU Method

In comparison to the other two methods, the RPU method resulted in more variability as to which items were detected based on the various study conditions. For instance, the hardest manipulated anchor item was either detected with perfect or zero power depending on the set of study conditions employed. In the instances for which it was zero, it was more likely that other items were detected. For the instances for which it was perfect, it was more likely that other items were not detected except for in a few cases as displayed in Tables 4.5 and 4.6. While more items were detected when there was a shift in the ability distribution between administrations, it is also true that the Type I error rate exceeded 0.05 and thus power must be interpreted with caution. These results are presented in the next section. Overall, this method appeared to be affected by shifts in the ability distribution, but not the correlation between the first administration's ability parameters. Additionally, the amount of IPD and number of items with simulated IPD did not produce consistent patterns.

4.2.1.2 Type I Error

The Type I error rate was reasonably controlled (i.e., less than 0.05) for most study conditions except for the following; (1) when the *b*-parameter plot method was employed under the condition for which administration 1 thetas are correlated at a 0.6 magnitude and there is no shift in the ability distribution between administrations with an *a*-shift of -0.3 and a *b*-shift of -0.5 for the condition where one item is simulated as aberrant and three items are simulated as aberrant and lie across the difficulty continuum, (2) when the *b*-parameter plot method was employed under the condition for which administration 1 thetas are correlated at a 0.6 magnitude and there is an ability shift between administrations with an *a*-shift of -0.7 and a *b*-shift of -0.5, and (3) for all conditions for which there is an ability distribution shift between administrations and the RPU method is employed. These results are presented in Tables 4.1 through 4.6 and in the figures presented in Appendix B. Further, very few control conditions (i.e., condition for which no items are simulated with drift) exhibit reasonable (i.e., less than 0.05) Type I error rates. For instance, looking at Table 4.1, where the delta plot method is employed, the average Type I error rate is 0.065, which exceeds the Type I error rate observed under conditions for which IPD was simulated. The exceptions include when no distributional shift is simulated and the delta or RPU detection methods are employed. Given the potential impact of the variation of detection rates and subsequent removal of problematic items from the anchor, ability parameter recovery is explored next.

4.2.2 Ability Parameter Recovery

For all study conditions employed, the RMSE fluctuated between approximately 0.3 and 0.4 theta points accounting for about a tenth of a standard deviation difference for accuracy. These results are displayed in Table 4.7 and are averaged across the 100 replications and across the entire ability distribution. Thus, no practical differences in the degree of recovery seem to exist. Appendix C displays plots for the average theta discrepancy of the ability estimate from the generating thetas for each tenth of theta point ranging from -2.5 to 2.5. While little variation existed between ability recovery rates, the magnitude of error is such that examinee classifications may be affected. These results are presented next.

4.2.3 Examinee Classification

Each examinee is classified based on the various calibrations, item detection methods employed, and resulting equating. From these classifications into proficiency categories and in comparison to the true classifications based on the cut scores (i.e., -0.75, 0.00, and 1.50) and generating thetas, the proportion of examinees accurately, under-, and over- classified can be determined. These proportions are presented in Tables 4.8 through 4.10, respectively. For ease of presentation, Figures 4.13 through 4.28 present the accurate classification rates pictorially and in comparison to the DTM thresholds calculated using no detection method as the point of comparison. Using the DTM criteria of 1.8% difference, very few practical differences exist in terms of the accurate classification of examinees in comparison to the case where no detection method is used for placing examinees into proficiency categories except for when a negatively skewed distribution is observed in the second administration (i.e., indicating the growth or distributional shift condition) when using the RPU method for item detection for the instances where three moderate to hard items and five spread items are simulated as aberrant. For such cases, an increase in the percent of examinees classified accurately is observed. An additional way to present these data appear in Appendix E where accurate, under-, and over- classification rates are presented in comparison to DTM thresholds calculated using the condition of no drift for comparison. Within these figures, similar trends emerge between the two different correlational conditions employed for administration 1. Additionally, it is only when growth is simulated between administrations that practical differences are observed such that the removal of flagged items results in an increase in the accuracy to which classification assignments are made. Further, using the RPU detection method resulted in fewer over-classifications and more accurate classifications. While no practical differences where observed between using the delta plot and IRT *b*-parameter plot methods, practical differences in classifications between these methods and the RPU method did emerge in the following instances where growth is simulated; (1) thetas for administration 1 are correlated 0.8, three items of moderate to hard difficulty are simulated as aberrant with a -0.7 *a*-shift and a -0.5 *b*-shift and (2) for both administration 1 conditions, five items are simulated as aberrant for all *a*-shift/*b*-shift combinations except the most mild (i.e., -0.3 *a*-shift and -0.5 *b*-shift) and ranged from 1.8 to 2.3 percent classification differences. The RPU method resulted in more under-classifications than any other method of detection employed with practically significant classification differences ranging from 1.8 to 3.4 percent. Likewise, fewer over-classifications were observed using the RPU method such that practically significant differences ranged from 1.8 to 5.5 percent. Additionally, most differences are observed within the over-classification table (Table 4.10) such that even within the condition of equivalent groups, practical differences are observed. Lastly, the tables in Appendix D present more detailed classification information within contingency tables explained in the methodology section and represented by Table 3.2. These tables collectively show that classification assignments are never off by more than one proficiency level under all conditions studied. In addition to exploring classification differences, it is also interesting to determine how these classification differences may translate into the amount of growth between administrations that would be reported under the varying conditions. The next section attempts to do this using the no drift condition as a proxy for the administration from which growth calculations are made.

4.2.4 Growth Expectations

It is expected that when no drift is simulated between administrations and growth is simulated between examinees, the result will be that examinees will score higher and thus may be placed in higher proficiency categories than their simulated analogs under the condition of no growth (i.e., the growth condition ability parameters were systematically manipulated based on the no growth ability parameters as described in Section 3.3.2). Thus, this condition is used to make all growth expectation conclusions for conditions where IPD is also simulated. Table 4.11 presents the observed percent of examinees that exhibit at least a one category increase in classification assignment based on the varied simulated conditions. A DTM is considered to exist when the difference in percentages between the control (i.e., no simulated drift) and the study conditions is greater than 1.8%. In such cases, the study condition is either assigned to a growth expectation status of “Under” or “Over.” An assignment of “Under” indicates that the amount of growth potentially reported would misrepresent truth by indicating that examinee performance had not improved as much as truth would indicate. An assignment of "Over" indicates that the amount of growth potentially reported would misrepresent truth by indicating that examinee performance had improved more than it really had. Finally, an assignment of "Meets" indicates that the amount of growth potentially reported would accurately represent truth by indicating that examinee performance had improved. As observed in Table 4.11, these designations do not differ between administration 1 conditions except in one case; when the 3 items that are spread across the difficultly continuum are simulated as aberrant and each have *a*-shifts of -0.3 and *b*-shifts of -0.8 and the *b*-parameter plot method is used for detecting aberrancy. In this case, when thetas are correlated 0.8 for the first administration, the resulting assessment of growth is over represented whereas it is not when thetas are correlated 0.6. Further, under all conditions, the RPU method misrepresents the amount of growth by indicating that examinee performance has not improved as much as truth would indicate. No other method under represents growth. Further, there are only three instances for which using either the delta plot or IRT *b*-parameter plots prevent an over representation of growth; (1) when one hard item is simulated as aberrant and there is an *a*-shift of -0.3 and a *b*-shift of -0.8, (2) when one hard item is simulated as aberrant and there is an *a*-shift of -0.7 and a *b*-shift of -0.8, and (3) when 5 items are simulated as aberrant, the delta plot method is employed, and there is an *a*-shift of -0.7 and a *b*-shift of -0.5. Otherwise, the use of no detection method is more likely to result in an over estimate of growth when items exhibiting IPD are uniformly easier and less discriminating in the second administration. The reverse is likely to be true when items are harder and more discriminating.

4.2.5 Summary of Simulation Study

In summary, the simulation study included five variables (i.e., administration 1 theta correlation condition, aberrant item scheme, degree of aberrancy combination, ability distribution shift condition, and aberrant item detection method) resulting in a total of 260 study conditions. Multiple interactions among conditions resulted, making it difficult to describe any main effects. Despite these multiple interactions, the effect of each of the study variables on the outcome variables (i.e., Type I error for aberrant item detection rates, power for aberrant item detection rates, ability parameter recovery, examinee classification accuracy, and growth outcomes) will be summarized in turn.

4.2.5.1 Administration 1 Theta Correlation

All else remaining constant, the difference in the degree of correlation between the administration 1 (i.e., 0.6 and 0.8) dimensions had minimal impact on all outcome variables assessed except in the following cases, (1) when employing the *b*-parameter plot method for the detection of aberrant items, Type I error was greater when a 0.6 correlation was simulated, (2) when growth was simulated between the two administrations such that the second set of examinees came from a negatively skewed distribution, power was more variable under a few conditions as specified in Section 4.2.1.1, (3) when dimensions were correlated 0.6, the *b*-parameter plot method was employed, and drift is simulated with an *a*-shift of 0.3 and *b*-shift of 0.8 for three items spread across the difficulty continuum, growth expectations were met. For this last observation when dimensions are correlated 0.8, an over prediction of growth results. The percent differences from what was expected were 1.6 and 2.1, respectively.

4.2.5.2 Aberrant Item Scheme

Four aberrant item schemes were simulated; (1) 1 difficult item drifted, (2) 1 easy, 1 moderate, and 1 difficult item drifted, (3) 2 moderate and 1 difficult item drifted, and (4) 2 easy, 2 moderate, and 1 difficult item drifted. Thus, the number and nature of the drifted items was explored. All else remaining constant, the differences observed between aberrant item schemes had no observable impact on Type I error rates for item detection yet great and variable impact on power rates for item detection such that power was most consistently the greatest when only one item was simulated as aberrant. A more detailed description of the exceptions and nature of power was discussed in Section 4.2.1.1. Lastly, ability parameter recovery remained unaffected and classification was such that a slight trend developed where as the number of aberrant items increased, the under-classification rates decreased.

4.2.5.3 Degree of Aberrancy Combination

Four aberrancy combinations were simulated where two shifts of the *a*-parameters (i.e., -0.3 & -0.7) and two shifts of the *b*-parameters (i.e., -0.5 & -0.8) were made uniformly across items simulated as aberrant and were fully crossed. Thus, the magnitude of drift was explored. All else remaining constant, the differences observed between magnitude combinations had minimal impact on Type I error rates for item detection. The exception occurs when the *b*-parameter plot method is employed and one difficult item or the three items spread condition was simulated. Under these specific conditions, the Type I error rates when the smaller magnitude (i.e., -0.5) of drift was simulated approached or exceeded the acceptable threshold of 0.05. Conversely, great and variable impact on power rates for item detection were observed as discussed in Section 4.2.1.1; yet, power very rarely reached acceptable levels of 0.85. Overall, there is no consistent pattern in relation to the drift magnitude combinations employed that can be described here. Ability parameter recovery and classification accuracy remained unaffected. The exception became noticeable (i.e., differences greater than 1.8%) only when growth (i.e., negatively skewed distribution) was simulated and five items were simulated with drift. In these cases and consistently across detection methods, more examinees were over-classified when an *a*-shift of -0.3 was simulated rather than -0.7. Lastly, the same pattern was observed for the error in the potentially reported growth percentage such that more error was present when an *a*-shift of -0.3 was simulated rather than -0.7, except when employing the RPU method where the trend was reversed (i.e., an *a*-shift of -0.7 resulted in more error in comparison to what was expected).

4.2.5.4 Ability Distribution Shift

Results were investigated under conditions where the two administration groups (i.e., first year and second year examinees) were equivalent (i.e., no distributional shift) and non-equivalent (i.e., distributional shift simulating growth) based on construct ability. Thus, the impact on outcome variables were explored under both non-growth and growth conditions. All else remaining constant, the differences observed between growth conditions had minimal impact on Type I error rates for item detection. The exception occurs when the RPU method is employed and will be discussed in Section 4.2.5.5. Great and variable impact on power rates for item detection were observed as discussed in Section 4.2.1.1; yet, power very rarely reached acceptable levels of 0.85. Overall, there is no consistent pattern in relation to the growth condition employed that can be described here. While the average ability parameter recovery rates were not significantly different, classification accuracy consistently declined when growth was simulated between administrations with fewer under-classifications occurring for the growth condition when employing delta and *b*-parameter plot detection methods and more over-classifications occurring when any of the detection methods were employed. Lastly, growth expectations were not explored for the no growth condition. Thus, no comparisons were made.

4.2.5.5 Aberrant Item Detection Method

Three aberrant item detection (i.e., item parameter drift) methods were implemented and include; (1) delta plot, (2) *b*-parameter plot, and (3) RPU methods. The Type I error rate was of the greatest concern for the RPU method when growth was simulated. Consistently across all conditions, it is more than twice the accepted level of 0.05. Thus, power must be interpreted with caution. Since the average power rate rarely reached acceptable levels, except in most cases where one item is simulated with IPD, and results are variable with no seeming patterns across studied variables, it is difficult to summarize differences in power other than the reported observations presented in Section 4.2.1.1. None of the methods employed can be singled out as adequate in this regard. Average ability parameter recovery rates and classification were not sensitive to the detection method employed, except for some conditions (i.e., as described in Section 4.2.3) when the RPU method is employed where both more accurate and more under-classifications result when a growth condition is simulated in comparison to the delta and *b*-parameter plotting methods. As a result of these differences, growth rates either met or exceeded the expected values for the delta and *b*-parameter plotting methods, while the RPU method under represents the expected growth.

Overall, too many interactions among conditions studied prevent clear interpretation of the results. Interpretations that can be summarized include the following; (1) the RPU method was sensitive to distributional shifts between administrations resulting in inflated Type I error rates, (2) no simulated drift resulted in inflated Type I error rates, (3) average IPD detection power was poor for most cases for which more than one item was simulated with IPD, (4) no differences in the average root mean squared error for ability parameter estimates were observed, (5) a shift in the ability distribution between administrations resulted in lower classification accuracy, (6) an increase in the number of aberrant items resulted in a decrease in the number of examinees under-classified, (7) when a growth condition was simulated, the RPU method resulted both in more accurate and more under-classifications than the delta and *b*-parameter plotting methods, (8) classification assignments were never off by more than one proficiency level, (9) the removal of aberrant items, regardless of whether it was correctly or incorrectly flagged, and regardless of the reason for the aberrancy, resulted in more accurate classifications and fewer over-classifications of examinees, (10) use of no detection method resulted in over estimates of growth when easier and less discriminating item appeared in the second administration, and (11) growth rates either met or exceeded the expected values for the delta and *b*-parameter plotting methods, while the RPU method under represented the expected growth. Next, the results from the empirical data analysis are presented and summarized.

4.3 Empirical Data Analysis

For this study, archived statewide science achievement assessment data were re-analyzed using the multiple IPD detection methods studied in the simulation. Based on item detection differences, different scaling constants were obtained for scaling ability parameter estimates between administrations and a set of fictitious cut scores (i.e., those used in the simulation study) were applied to the resulting parameter estimates to classify examinees into proficiency categories. Fictitious, rather than operational cuts were employed to be consistent with the simulation study design and analysis. Calibrations were performed separately for each administration and the number of items that required the *c*-parameter to be fixed due to non-convergence is presented in Table A.8 of Appendix A. IPD analysis summaries are followed by the summary of the resulting scaling constants and proficiency classification differences.

4.3.1 Item Parameter Drift Analysis

Each of the IPD detection methods were applied to the grade 4, 8, and 11 science assessments. Flagged items are reported as a ‘1’ in the ‘Flag’ columns of Tables 4.12 through 4.14 for the delta analysis, Tables 4.15 through 4.17 for the IRT *b*-parameter plot analysis, and Tables 4.18 through 4.20 for the RPU method analysis. All methods flagged different items and when using two standard deviations rather than three for the delta method, more items were flagged. In fact, when using three standard deviations as the criterion, no items were flagged. Item 26 for grade 4 was flagged using both the delta plot method with two standard deviations as the criterion and using the *b*-parameter plot method. Likewise, item 38 for grade 8 was flagged using both the delta plot method with two standard deviations as the criterion and using the RPU method. Further, in grade 11, item 37 was flagged using both the *b*-parameter plot method and the RPU method. Otherwise, only one method flagged any particular items as observed in Tables 4.12 through 4.20. Figures 4.29 through 4.42 display the ICC and ICCC plots for all flagged items for each administration to help visually determine whether the item should have been flagged. Noticeable differences are observed in all the plots except for item 12 for grade 4, items 17 and 22 for grade 8, and item 14 for grade 11 for which all were flagged using the *b*-parameter plot method. The only item for which the *b*-parameter plot method serves as the only method that detects a questionable item occurs for item 3 in grade 11. Regardless of the subjective analysis of the ICC and ICCC plots, flagged items were removed and the summary of the equating as a result of item removal decisions is presented next.

4.3.2 Equating Summary

Based on the IPD analyses, different equating analyses were conducted. Table 4.21 displays the resulting scaling constants for each of the grade levels. These scaling constants were applied to the ability estimates for the second years' administration and resulting estimates were then classified into four proficiency categories as described next.

4.3.3 Proficiency Classification

Examinees were classified into proficiency categories based on the cut scores used in the simulation study (i.e., -0.75, 0.00, and 1.50). Tables 4.22 through 4.24 present the proportion of examinees assigned to each of the categories based on the detection method employed and the resulting equating constants. In practice, a score that would be considered as meeting AYP would be one that is classified into either category 3 or 4.[5]

[5] The cut scores implemented in this study are not the actual cut scores applied in practice for this assessment.

Thus, the ‘MEETS AYP’ column indicates the proportion of examinees that would contribute to meeting AYP. The last two columns represent the difference between the proportions of examinees contributing in comparison to the first administration and the difference between the proportions of examinees contributing in comparison to the standard method for assessing for IPD, respectively. Using the operational approach (i.e., delta plot method with a criterion of three standard deviations from the line of best fit) for IPD detection, growth is observed in grades 4 and 11 such that 2.9% more and 3.1% more examinees are contributing to meeting AYP, respectively. Using the DTM criterion of 1.8% difference, it is also observed that when using any other method for item detection, the percent change of examinees contributing to meeting AYP between administrations would not significantly change for grade 4. For grade 8, there was no practical change in the proportion of examinees contributing to meeting AYP. However, if using the RPU method for detecting aberrant items, a growth of 2.9% would be observed. Likewise, while the other methods would not indicate practically significant growth between administrations, practically significant differences do exist between using the operational delta plot method with a criterion of 3 SDs and all other detection methods as observed in the last column of Table 4.23. Lastly, for grade 11, there is a practical difference in the change when the RPU method is employed such that 2.1% fewer examinees would be considered as contributing toward meeting AYP. Here, it is also interesting to note that the items the RPU method flagged were both polytomous items that were easier in the second administration according to their global *b*-parameter estimates. Thus, removing these items from equating would produce substantial differences as they also contribute four points each to the total number of points used for linking forms.

4.3.4 Summary of Empirical Data Analysis

The empirical data analysis consisted of re-analyzing archived statewide science achievement assessment data by applying the multiple IPD detection methods studied in the simulation. Doing so resulted in all methods flagging different items. The delta method for which the criterion of three standard deviations is used resulted in flagging no items as aberrant; and thus, growth was observed in grades 4 and 11 such that 2.9% more and 3.1% more examinees were contributing to meeting AYP, respectively. No growth was observed for grade 8 examinees. Of all items flagged by any given alternative method and for each of the three assessments, only one item was flagged by more than one detection method. Otherwise, no more than three items were flagged for any given method and across all methods a total of four or five items were flagged. Especially worthy of note, only polytomous items were flagged when using the RPU procedures for detecting aberrant items. It was also observed that when using any other method for item detection, the percent change of examinees contributing to meeting AYP between administrations would not significantly change for grade 4. For grade 11, there is a practical difference in the change when the RPU method is employed such that 2.1% fewer examinees would be considered as contributing toward meeting AYP. For grade 8, all methods except for the RPU method resulted in the observation of no growth between administrations. However, when using the RPU method for detecting aberrant items, a growth of 2.9% would have been observed.

The results from both the simulation and empirical data analyses are further discussed in Chapter 5. It is here that possible explanations for major findings, connections between simulation and empirical studies, limitations of this work, further research directions, and final conclusions and recommendations for practice are presented.

Table 4.1. Power & Type I Error for the Delta Plot Method of Item Parameter Drift Detection, Administration 1 Correlated Thetas by 0.8, by Condition

	Condition	Amount of IPD				Power			Type I
Distribution	Description	*a*	*b*	Item 1	Item 2	Item 3	Item 4	Item 5	Error
Normal (No Shift)	No IPD	None	None	▨	▨	▨	▨	▨	0.034
	1 Item, Hard	-0.3	-0.5	1.000					0.000
		-0.7	-0.5	1.000					0.000
		-0.3	-0.8	1.000					0.000
		-0.7	-0.8	1.000					0.000
	3 Items, Spread	-0.3	-0.5	1.000	0.000	0.000			0.000
		-0.7	-0.5	1.000	0.060	0.000			0.000
		-0.3	-0.8	1.000	0.640	1.000			0.000
		-0.7	-0.8	1.000	0.000	0.000			0.000
	3 Items, Moderate	-0.3	-0.5	1.000		0.000	0.000		0.000
		-0.7	-0.5	1.000		0.000	0.000		0.000
		-0.3	-0.8	1.000		0.000	0.000		0.000
		-0.7	-0.8	1.000		0.000	0.000		0.000
	5 Items, Spread	-0.3	-0.5	1.000	0.000	0.000	0.000	0.000	0.000
		-0.7	-0.5	1.000	0.130	0.000	0.000	0.000	0.000
		-0.3	-0.8	1.000	0.150	0.000	0.000	0.000	0.000
		-0.7	-0.8	1.000	0.000	0.000	0.000	0.000	0.000
Negatively Skewed (Shift)	No IPD	None	None	▨	▨	▨	▨	▨	0.065
	1 Item, Hard	-0.3	-0.5	1.000					0.000
		-0.7	-0.5	1.000					0.003
		-0.3	-0.8	1.000					0.000
		-0.7	-0.8	1.000					0.001
	3 Items, Spread	-0.3	-0.5	1.000	0.000	0.000			0.000
		-0.7	-0.5	1.000	0.000	0.000			0.007
		-0.3	-0.8	1.000	0.000	0.830			0.000
		-0.7	-0.8	1.000	0.060	0.650			0.001
	3 Items, Moderate	-0.3	-0.5	1.000		0.000	0.000		0.004
		-0.7	-0.5	1.000		0.000	0.000		0.000
		-0.3	-0.8	1.000		0.910	0.000		0.000
		-0.7	-0.8	1.000		0.000	0.000		0.000
	5 Items, Spread	-0.3	-0.5	1.000	0.000	0.000	0.280	0.000	0.002
		-0.7	-0.5	1.000	1.000	0.000	0.000	0.000	0.000
		-0.3	-0.8	1.000	0.000	0.020	1.000	0.000	0.000
		-0.7	-0.8	1.000	1.000	0.000	0.000	0.000	0.000

Table 4.2. Power & Type I Error for the Delta Plot Method of Item Parameter Drift Detection, Administration 1 Correlated Thetas by 0.6, by Condition

	Condition	Amount of IPD		Power					Type I
Distribution	Description	*a*	*b*	Item 1	Item 2	Item 3	Item 4	Item 5	Error
Normal (No Shift)	No IPD	None	None	▨	▨	▨	▨	▨	0.034
	1 Item, Hard	-0.3	-0.5	1.000					0.000
		-0.7	-0.5	1.000					0.000
		-0.3	-0.8	1.000					0.000
		-0.7	-0.8	1.000					0.000
	3 Items, Spread	-0.3	-0.5	1.000	0.000	0.000			0.000
		-0.7	-0.5	1.000	0.030	0.000			0.000
		-0.3	-0.8	1.000	0.780	1.000			0.000
		-0.7	-0.8	1.000	0.000	0.000			0.000
	3 Items, Moderate	-0.3	-0.5	1.000		0.000	0.000		0.000
		-0.7	-0.5	1.000		0.000	0.000		0.000
		-0.3	-0.8	1.000		0.000	0.000		0.000
		-0.7	-0.8	1.000		0.000	0.000		0.000
	5 Items, Spread	-0.3	-0.5	1.000	0.000	0.000	0.000	0.000	0.000
		-0.7	-0.5	1.000	0.060	0.000	0.000	0.000	0.000
		-0.3	-0.8	1.000	0.120	0.000	0.000	0.000	0.000
		-0.7	-0.8	1.000	0.000	0.000	0.000	0.000	0.000
Negatively Skewed (Shift)	No IPD	None	None	▨	▨	▨	▨	▨	0.063
	1 Item, Hard	-0.3	-0.5	1.000					0.000
		-0.7	-0.5	1.000					0.003
		-0.3	-0.8	1.000					0.000
		-0.7	-0.8	1.000					0.001
	3 Items, Spread	-0.3	-0.5	1.000	0.000	0.000			0.000
		-0.7	-0.5	1.000	0.000	0.000			0.007
		-0.3	-0.8	1.000	0.000	0.800			0.000
		-0.7	-0.8	1.000	0.090	0.610			0.001
	3 Items, Moderate	-0.3	-0.5	1.000		0.000	0.000		0.004
		-0.7	-0.5	1.000		0.000	0.000		0.000
		-0.3	-0.8	1.000		0.900	0.000		0.000
		-0.7	-0.8	1.000		0.000	0.000		0.000
	5 Items, Spread	-0.3	-0.5	1.000	0.000	0.000	0.000	0.490	0.002
		-0.7	-0.5	1.000	1.000	0.000	0.000	0.000	0.000
		-0.3	-0.8	1.000	0.000	0.000	0.000	1.000	0.000
		-0.7	-0.8	1.000	1.000	0.000	0.000	0.000	0.000

Table 4.3. Power & Type I Error for the *b* Plot Method of Item Parameter Drift Detection Administration 1 Correlated Thetas by 0.8, by Condition

	Condition	Amount of IPD		Power					Type I
Distribution	Description	*a*	*b*	Item 1	Item 2	Item 3	Item 4	Item 5	Error
Normal (No Shift)	No IPD	None	None	▨	▨	▨	▨	▨	0.065
	1 Item, Hard	-0.3	-0.5	1.000					0.036
		-0.7	-0.5	1.000					0.007
		-0.3	-0.8	1.000					0.030
		-0.7	-0.8	1.000					0.007
	3 Items, Spread	-0.3	-0.5	1.000	0.010	0.060			0.034
		-0.7	-0.5	0.820	0.740	0.010			0.004
		-0.3	-0.8	1.000	0.170	0.390			0.024
		-0.7	-0.8	0.770	0.820	0.000			0.004
	3 Items, Moderate	-0.3	-0.5	0.810		0.000	0.000		0.027
		-0.7	-0.5	0.940		0.020	0.410		0.012
		-0.3	-0.8	0.950		0.010	0.490		0.013
		-0.7	-0.8	0.940		0.040	0.290		0.015
	5 Items, Spread	-0.3	-0.5	0.940	0.000	0.060	0.420	0.000	0.014
		-0.7	-0.5	0.820	0.730	0.000	0.200	0.000	0.003
		-0.3	-0.8	0.830	0.770	0.000	0.210	0.000	0.004
		-0.7	-0.8	0.770	0.800	0.000	0.520	0.000	0.002
Negatively Skewed (Shift)	No IPD	None	None	▨	▨	▨	▨	▨	0.069
	1 Item, Hard	-0.3	-0.5	1.000					0.027
		-0.7	-0.5	0.990					0.034
		-0.3	-0.8	1.000					0.017
		-0.7	-0.8	1.000					0.025
	3 Items, Spread	-0.3	-0.5	1.000	0.020	0.050			0.024
		-0.7	-0.5	0.980	0.040	0.070			0.032
		-0.3	-0.8	1.000	0.280	0.630			0.010
		-0.7	-0.8	1.000	0.190	0.360			0.021
	3 Items, Moderate	-0.3	-0.5	1.000		0.060	0.130		0.022
		-0.7	-0.5	0.980		0.040	0.470		0.003
		-0.3	-0.8	1.000		0.570	0.590		0.009
		-0.7	-0.8	0.980		0.070	0.330		0.005
	5 Items, Spread	-0.3	-0.5	1.000	0.010	0.030	0.080	0.080	0.021
		-0.7	-0.5	0.900	0.760	0.010	0.290	0.010	0.001
		-0.3	-0.8	1.000	0.100	0.340	0.420	0.360	0.004
		-0.7	-0.8	0.820	0.850	0.010	0.600	0.010	0.000

Table 4.4. Power & Type I Error for the *b* Plot Method of Item Parameter Drift Detection, Administration 1 Correlated Thetas by 0.6, by Condition

Distribution	Condition Description	Amount of IPD		Power					Type I Error
		a	*b*	Item 1	Item 2	Item 3	Item 4	Item 5	
Normal (No Shift)	No IPD	None	None	▨	▨	▨	▨	▨	0.057
	1 Item, Hard	-0.3	-0.5	0.570					0.052
		-0.7	-0.5	1.000					0.015
		-0.3	-0.8	0.960					0.044
		-0.7	-0.8	1.000					0.015
	3 Items, Spread	-0.3	-0.5	0.510	0.020	0.080			0.052
		-0.7	-0.5	0.870	0.720	0.010			0.005
		-0.3	-0.8	0.930	0.380	0.450			0.034
		-0.7	-0.8	0.830	0.800	0.000			0.004
	3 Items, Moderate	-0.3	-0.5	0.850		0.000	0.000		0.027
		-0.7	-0.5	1.000		0.030	0.440		0.015
		-0.3	-0.8	1.000		0.020	0.500		0.015
		-0.7	-0.8	1.000		0.020	0.370		0.017
	5 Items, Spread	-0.3	-0.5	1.000	0.000	0.050	0.400	0.000	0.017
		-0.7	-0.5	0.860	0.720	0.010	0.260	0.000	0.004
		-0.3	-0.8	0.910	0.760	0.000	0.260	0.000	0.005
		-0.7	-0.8	0.820	0.800	0.000	0.500	0.000	0.003
Negatively Skewed (Shift)	No IPD	None	None	▨	▨	▨	▨	▨	0.059
	1 Item, Hard	-0.3	-0.5	0.890					0.049
		-0.7	-0.5	0.630					0.050
		-0.3	-0.8	1.000					0.039
		-0.7	-0.8	0.910					0.043
	3 Items, Spread	-0.3	-0.5	0.830	0.060	0.080			0.046
		-0.7	-0.5	0.560	0.070	0.070			0.049
		-0.3	-0.8	1.000	0.580	0.810			0.023
		-0.7	-0.8	0.830	0.360	0.400			0.032
	3 Items, Moderate	-0.3	-0.5	0.850		0.070	0.270		0.047
		-0.7	-0.5	0.990		0.050	0.010		0.004
		-0.3	-0.8	1.000		0.820	0.100		0.023
		-0.7	-0.8	0.990		0.060	0.020		0.005
	5 Items, Spread	-0.3	-0.5	0.760	0.020	0.030	0.220	0.260	0.041
		-0.7	-0.5	0.920	0.760	0.010	0.010	0.010	0.001
		-0.3	-0.8	0.950	0.250	0.480	0.040	0.830	0.016
		-0.7	-0.8	0.850	0.870	0.010	0.000	0.010	0.000

Table 4.5. Power & Type I Error for the RPU Method of Item Parameter Drift Detection, Administration 1 Correlated Thetas by 0.8, by Condition

	Condition	Amount of IPD		Power					Type I
Distribution	Description	*a*	*b*	Item 1	Item 2	Item 3	Item 4	Item 5	Error
Normal (No Shift)	No IPD	None	None	▨	▨	▨	▨	▨	0.000
	1 Item, Hard	-0.3	-0.5	0.000					0.000
		-0.7	-0.5	1.000					0.000
		-0.3	-0.8	0.000					0.000
		-0.7	-0.8	1.000					0.000
	3 Items, Spread	-0.3	-0.5	0.000	0.000	0.000			0.000
		-0.7	-0.5	1.000	0.080	0.000			0.000
		-0.3	-0.8	0.000	0.890	0.990			0.000
		-0.7	-0.8	1.000	0.030	0.000			0.004
	3 Items, Moderate	-0.3	-0.5	1.000		0.000	0.000		0.006
		-0.7	-0.5	1.000		0.000	0.000		0.000
		-0.3	-0.8	1.000		0.000	0.000		0.000
		-0.7	-0.8	1.000		0.000	0.000		0.000
	5 Items, Spread	-0.3	-0.5	1.000	0.000	0.020	0.000	0.000	0.000
		-0.7	-0.5	1.000	0.050	0.000	0.000	0.000	0.000
		-0.3	-0.8	1.000	0.030	0.000	0.000	0.000	0.000
		-0.7	-0.8	1.000	0.020	0.000	0.850	0.000	0.003
Negatively Skewed (Shift)	No IPD	None	None	▨	▨	▨	▨	▨	0.137
	1 Item, Hard	-0.3	-0.5	0.000					0.141
		-0.7	-0.5	0.000					0.128
		-0.3	-0.8	0.000					0.141
		-0.7	-0.8	0.000					0.128
	3 Items, Spread	-0.3	-0.5	0.000	1.000	0.870			0.129
		-0.7	-0.5	0.000	1.000	0.640			0.118
		-0.3	-0.8	0.000	1.000	1.000			0.129
		-0.7	-0.8	0.000	1.000	1.000			0.118
	3 Items, Moderate	-0.3	-0.5	0.000		0.860	0.000		0.150
		-0.7	-0.5	1.000		0.070	0.000		0.155
		-0.3	-0.8	0.000		1.000	0.990		0.148
		-0.7	-0.8	1.000		0.960	0.350		0.155
	5 Items, Spread	-0.3	-0.5	0.000	1.000	0.880	0.000	1.000	0.112
		-0.7	-0.5	1.000	0.290	0.070	0.000	0.000	0.115
		-0.3	-0.8	0.000	1.000	1.000	0.990	1.000	0.112
		-0.7	-0.8	1.000	0.150	0.070	1.000	0.510	0.143

Table 4.6. Power & Type I Error for the RPU Method of Item Parameter Drift Detection, Administration 1 Correlated Thetas by 0.6, by Condition

Distribution	Condition Description	Amount of IPD		Power					Type I Error
		a	*b*	Item 1	Item 2	Item 3	Item 4	Item 5	
Normal (No Shift)	No IPD	None	None	▨	▨	▨	▨	▨	0.000
	1 Item, Hard	-0.3	-0.5	0.000					0.000
		-0.7	-0.5	1.000					0.000
		-0.3	-0.8	0.000					0.000
		-0.7	-0.8	1.000					0.000
	3 Items, Spread	-0.3	-0.5	0.000	0.000	0.000			0.000
		-0.7	-0.5	1.000	0.080	0.000			0.000
		-0.3	-0.8	0.000	0.910	0.990			0.000
		-0.7	-0.8	1.000	0.040	0.000			0.004
	3 Items, Moderate	-0.3	-0.5	1.000		0.000	0.000		0.006
		-0.7	-0.5	1.000		0.000	0.000		0.000
		-0.3	-0.8	1.000		0.000	0.000		0.000
		-0.7	-0.8	1.000		0.020	0.000		0.000
	5 Items, Spread	-0.3	-0.5	1.000	0.000	0.050	0.000	0.000	0.000
		-0.7	-0.5	1.000	0.050	0.000	0.000	0.000	0.000
		-0.3	-0.8	1.000	0.030	0.000	0.000	0.000	0.000
		-0.7	-0.8	1.000	0.020	0.000	0.880	0.000	0.003
Negatively Skewed (Shift)	No IPD	None	None	▨	▨	▨	▨	▨	0.146
	1 Item, Hard	-0.3	-0.5	0.000					0.150
		-0.7	-0.5	0.000					0.140
		-0.3	-0.8	0.000					0.150
		-0.7	-0.8	0.000					0.140
	3 Items, Spread	-0.3	-0.5	0.000	1.000	1.000			0.133
		-0.7	-0.5	0.000	1.000	0.960			0.125
		-0.3	-0.8	0.000	1.000	1.000			0.133
		-0.7	-0.8	0.000	1.000	1.000			0.125
	3 Items, Moderate	-0.3	-0.5	0.000		1.000	0.000		0.159
		-0.7	-0.5	1.000		0.250	0.100		0.161
		-0.3	-0.8	0.000		1.000	1.000		0.158
		-0.7	-0.8	1.000		1.000	0.890		0.160
	5 Items, Spread	-0.3	-0.5	0.000	1.000	1.000	0.000	1.000	0.112
		-0.7	-0.5	1.000	0.290	0.260	0.100	0.000	0.115
		-0.3	-0.8	0.000	1.000	1.000	1.000	1.000	0.112
		-0.7	-0.8	1.000	0.150	0.270	1.000	0.850	0.143

Table 4.7. Root Mean Squared Error Mean and Standard Deviation (SD) for Theta Estimates Obtained for No Item Parameter Drift & Varying Conditions of Item Parameter Drift and Detection, Across 100 Replications

			Normal (No Shift) Distribution							
			None		Delta		IRT		RPU	
Administration 1 Condition	Condition Description	Degree of Aberrancy (*a/b-shifts*)	Mean	SD	Mean	SD	Mean	SD	Mean	SD
1	1 Item, Hard	0.3 / 0.5	0.39	0.01	0.39	0.01	0.38	0.01	0.39	0.01
		0.7 / 0.5	0.39	0.01	0.39	0.01	0.38	0.01	0.39	0.01
		0.3 / 0.8	0.39	0.01	0.39	0.01	0.38	0.01	0.39	0.01
		0.7 / 0.8	0.39	0.01	0.39	0.01	0.38	0.01	0.39	0.01
	3 Items, Spread	0.3 / 0.5	0.38	0.01	0.37	0.01	0.37	0.01	0.38	0.01
		0.7 / 0.5	0.36	0.01	0.35	0.01	0.37	0.01	0.35	0.01
		0.3 / 0.8	0.38	0.01	0.38	0.01	0.37	0.01	0.38	0.01
		0.7 / 0.8	0.36	0.01	0.35	0.01	0.37	0.01	0.35	0.01
	3 Items, Moderate	0.3 / 0.5	0.36	0.01	0.35	0.01	0.37	0.01	0.35	0.01
		0.7 / 0.5	0.37	0.01	0.36	0.01	0.36	0.01	0.36	0.01
		0.3 / 0.8	0.37	0.01	0.36	0.01	0.36	0.01	0.36	0.01
		0.7 / 0.8	0.37	0.01	0.36	0.01	0.36	0.01	0.36	0.01
	5 Items, Spread	0.3 / 0.5	0.37	0.01	0.36	0.01	0.36	0.01	0.36	0.01
		0.7 / 0.5	0.34	0.01	0.33	0.01	0.34	0.01	0.33	0.01
		0.3 / 0.8	0.34	0.01	0.33	0.01	0.34	0.01	0.33	0.01
		0.7 / 0.8	0.34	0.01	0.33	0.00	0.35	0.01	0.33	0.01
2	1 Item, Hard	0.3 / 0.5	0.38	0.01	0.38	0.01	0.37	0.01	0.38	0.01
		0.7 / 0.5	0.38	0.01	0.38	0.01	0.37	0.01	0.38	0.01
		0.3 / 0.8	0.38	0.01	0.38	0.01	0.37	0.01	0.38	0.01
		0.7 / 0.8	0.38	0.01	0.38	0.01	0.37	0.01	0.38	0.01
	3 Items, Spread	0.3 / 0.5	0.37	0.01	0.37	0.01	0.36	0.01	0.37	0.01
		0.7 / 0.5	0.35	0.01	0.35	0.01	0.36	0.01	0.35	0.01
		0.3 / 0.8	0.37	0.01	0.37	0.01	0.37	0.01	0.38	0.01
		0.7 / 0.8	0.35	0.01	0.34	0.01	0.36	0.01	0.34	0.01
	3 Items, Moderate	0.3 / 0.5	0.35	0.01	0.34	0.01	0.36	0.01	0.34	0.01
		0.7 / 0.5	0.36	0.01	0.35	0.01	0.36	0.01	0.35	0.01
		0.3 / 0.8	0.36	0.01	0.35	0.01	0.36	0.01	0.35	0.01
		0.7 / 0.8	0.36	0.01	0.36	0.01	0.36	0.01	0.36	0.01
	5 Items, Spread	0.3 / 0.5	0.36	0.01	0.36	0.01	0.36	0.01	0.36	0.01
		0.7 / 0.5	0.33	0.01	0.33	0.01	0.34	0.01	0.33	0.01
		0.3 / 0.8	0.33	0.01	0.33	0.01	0.34	0.01	0.33	0.01
		0.7 / 0.8	0.34	0.01	0.33	0.00	0.34	0.01	0.33	0.01

Continued, next page.

Table 4.7., cont'd.:

Table 4.7. Root Mean Squared Error Mean and Standard Deviation (SD) for Theta Estimates Obtained for No Item Parameter Drift & Varying Conditions of Item Parameter Drift and Detection, Across 100 Replications

			Negatively Skewed (Shift) Distribution							
			None		Delta		IRT		RPU	
Administration 1 Condition	Condition Description	Degree of Aberrancy (*a/b-shifts*)	Mean	SD	Mean	SD	Mean	SD	Mean	SD
1	1 Item, Hard	0.3 / 0.5	0.37	0.00	0.36	0.00	0.36	0.00	0.35	0.01
		0.7 / 0.5	0.36	0.01	0.36	0.01	0.36	0.01	0.36	0.01
		0.3 / 0.8	0.37	0.00	0.36	0.00	0.36	0.00	0.35	0.01
		0.7 / 0.8	0.37	0.01	0.36	0.01	0.36	0.01	0.36	0.01
	3 Items, Spread	0.3 / 0.5	0.36	0.00	0.36	0.00	0.35	0.00	0.35	0.01
		0.7 / 0.5	0.36	0.01	0.36	0.01	0.35	0.01	0.35	0.01
		0.3 / 0.8	0.37	0.00	0.35	0.00	0.35	0.01	0.35	0.01
		0.7 / 0.8	0.37	0.01	0.35	0.01	0.35	0.01	0.35	0.01
	3 Items, Moderate	0.3 / 0.5	0.37	0.00	0.36	0.00	0.36	0.00	0.35	0.01
		0.7 / 0.5	0.36	0.01	0.35	0.00	0.35	0.00	0.33	0.01
		0.3 / 0.8	0.37	0.00	0.36	0.00	0.36	0.01	0.35	0.01
		0.7 / 0.8	0.36	0.01	0.35	0.00	0.35	0.00	0.34	0.01
	5 Items, Spread	0.3 / 0.5	0.36	0.00	0.35	0.00	0.35	0.00	0.35	0.01
		0.7 / 0.5	0.33	0.01	0.33	0.00	0.33	0.01	0.31	0.01
		0.3 / 0.8	0.37	0.00	0.36	0.00	0.36	0.01	0.35	0.01
		0.7 / 0.8	0.34	0.01	0.34	0.00	0.34	0.01	0.31	0.01
2	1 Item, Hard	0.3 / 0.5	0.36	0.00	0.35	0.00	0.35	0.00	0.34	0.01
		0.7 / 0.5	0.36	0.01	0.35	0.01	0.35	0.01	0.35	0.01
		0.3 / 0.8	0.36	0.00	0.35	0.00	0.35	0.00	0.35	0.01
		0.7 / 0.8	0.36	0.01	0.35	0.01	0.35	0.01	0.35	0.01
	3 Items, Spread	0.3 / 0.5	0.35	0.00	0.35	0.00	0.35	0.00	0.34	0.01
		0.7 / 0.5	0.35	0.01	0.35	0.01	0.35	0.01	0.35	0.01
		0.3 / 0.8	0.36	0.00	0.35	0.00	0.35	0.00	0.34	0.01
		0.7 / 0.8	0.36	0.01	0.35	0.01	0.35	0.01	0.35	0.01
	3 Items, Moderate	0.3 / 0.5	0.36	0.00	0.36	0.00	0.35	0.00	0.34	0.01
		0.7 / 0.5	0.35	0.01	0.34	0.00	0.34	0.01	0.32	0.01
		0.3 / 0.8	0.37	0.00	0.36	0.00	0.35	0.00	0.34	0.01
		0.7 / 0.8	0.35	0.01	0.34	0.00	0.34	0.00	0.34	0.01
	5 Items, Spread	0.3 / 0.5	0.35	0.00	0.35	0.00	0.35	0.00	0.34	0.01
		0.7 / 0.5	0.32	0.01	0.33	0.00	0.33	0.01	0.30	0.01
		0.3 / 0.8	0.36	0.00	0.35	0.00	0.35	0.00	0.34	0.01
		0.7 / 0.8	0.33	0.01	0.33	0.00	0.33	0.01	0.31	0.01

Table 4.8. Proportion of Examinees Accurately Classified, by Condition

			Normal (No Shift) Distribution			
			Degree of Aberrancy (a-shift/b-shift)			
Administration 1 Condition	Condition Description	Detection Method	0.3 / 0.5	0.7 / 0.5	0.3 / 0.8	0.7 / 0.8
1	1 Item, Hard	None	0.785	0.781	0.784	0.781
		Delta	0.787	0.786	0.787	0.786
		IRT	0.788	0.787	0.788	0.787
		RPU	0.785	0.786	0.784	0.786
	3 Items, Spread	None	0.787	0.792	0.784	0.791
		Delta	0.788	0.796	0.789	0.796
		IRT	0.790	0.791	0.789	0.791
		RPU	0.787	0.796	0.786	0.796
	3 Items, Moderate	None	0.791	0.785	0.785	0.784
		Delta	0.797	0.792	0.792	0.791
		IRT	0.791	0.791	0.791	0.791
		RPU	0.797	0.792	0.792	0.791
	5 Items, Spread	None	0.784	0.796	0.796	0.792
		Delta	0.791	0.801	0.802	0.800
		IRT	0.791	0.797	0.797	0.795
		RPU	0.791	0.802	0.802	0.801
2	1 Item, Hard	None	0.789	0.784	0.788	0.784
		Delta	0.789	0.789	0.789	0.789
		IRT	0.791	0.790	0.791	0.790
		RPU	0.789	0.789	0.788	0.789
	3 Items, Spread	None	0.790	0.794	0.788	0.794
		Delta	0.791	0.799	0.790	0.799
		IRT	0.792	0.794	0.791	0.793
		RPU	0.790	0.798	0.789	0.799
	3 Items, Moderate	None	0.794	0.787	0.788	0.787
		Delta	0.799	0.794	0.794	0.793
		IRT	0.793	0.793	0.793	0.793
		RPU	0.799	0.794	0.794	0.793
	5 Items, Spread	None	0.787	0.798	0.798	0.794
		Delta	0.793	0.804	0.804	0.802
		IRT	0.793	0.799	0.799	0.797
		RPU	0.793	0.804	0.804	0.802

Continued, next page.

Table 4.8., cont'd.:

Table 4.8. Proportion of Examinees Accurately Classified, by Condition

			Negatively Skewed (Shift) Distribution *Degree of Aberrancy (a-shift/b-shift)*			
Administration 1 Condition	Condition Description	Detection Method	0.3 / 0.5	0.7 / 0.5	0.3 / 0.8	0.7 / 0.8
1	1 Item, Hard	None	0.761	0.761	0.759	0.759
		Delta	0.763	0.763	0.763	0.763
		IRT	0.764	0.765	0.764	0.764
		RPU	0.773	0.771	0.772	0.770
	3 Items, Spread	None	0.760	0.761	0.755	0.756
		Delta	0.762	0.763	0.763	0.763
		IRT	0.763	0.764	0.763	0.763
		RPU	0.775	0.772	0.773	0.772
	3 Items, Moderate	None	0.756	0.750	0.750	0.749
		Delta	0.758	0.765	0.759	0.763
		IRT	0.759	0.764	0.760	0.763
		RPU	0.774	0.782	0.774	0.779
	5 Items, Spread	None	0.756	0.765	0.747	0.759
		Delta	0.760	0.772	0.756	0.767
		IRT	0.760	0.772	0.757	0.768
		RPU	0.774	0.790	0.774	0.790
2	1 Item, Hard	None	0.764	0.765	0.763	0.763
		Delta	0.766	0.766	0.766	0.766
		IRT	0.768	0.768	0.768	0.768
		RPU	0.777	0.775	0.776	0.774
	3 Items, Spread	None	0.764	0.764	0.758	0.759
		Delta	0.765	0.765	0.765	0.765
		IRT	0.767	0.767	0.767	0.765
		RPU	0.777	0.776	0.776	0.774
	3 Items, Moderate	None	0.759	0.754	0.754	0.752
		Delta	0.761	0.767	0.761	0.766
		IRT	0.763	0.767	0.763	0.766
		RPU	0.777	0.783	0.776	0.778
	5 Items, Spread	None	0.760	0.767	0.750	0.761
		Delta	0.763	0.775	0.758	0.769
		IRT	0.763	0.774	0.762	0.769
		RPU	0.776	0.792	0.776	0.789

Table 4.9. Proportion of Examinees Under-classified, by Condition

			Normal (No Shift) Distribution			
			Degree of Aberrancy (a-shift/b-shift)			
Administration 1 Condition	Condition Description	Detection Method	0.3 / 0.5	0.7 / 0.5	0.3 / 0.8	0.7 / 0.8
1	1 Item, Hard	None	0.083	0.066	0.082	0.066
		Delta	0.084	0.085	0.084	0.085
		IRT	0.084	0.084	0.084	0.084
		RPU	0.083	0.085	0.082	0.085
	3 Items, Spread	None	0.076	0.065	0.071	0.063
		Delta	0.077	0.080	0.082	0.078
		IRT	0.077	0.078	0.075	0.075
		RPU	0.076	0.081	0.081	0.079
	3 Items, Moderate	None	0.064	0.059	0.059	0.058
		Delta	0.079	0.073	0.073	0.071
		IRT	0.076	0.074	0.074	0.072
		RPU	0.080	0.073	0.073	0.071
	5 Items, Spread	None	0.058	0.059	0.060	0.054
		Delta	0.072	0.074	0.075	0.068
		IRT	0.073	0.072	0.073	0.068
		RPU	0.072	0.074	0.075	0.073
2	1 Item, Hard	None	0.082	0.065	0.080	0.065
		Delta	0.083	0.083	0.083	0.083
		IRT	0.082	0.083	0.083	0.083
		RPU	0.082	0.083	0.080	0.083
	3 Items, Spread	None	0.075	0.064	0.069	0.062
		Delta	0.076	0.079	0.082	0.077
		IRT	0.075	0.077	0.076	0.074
		RPU	0.075	0.080	0.080	0.078
	3 Items, Moderate	None	0.063	0.057	0.058	0.056
		Delta	0.078	0.072	0.072	0.070
		IRT	0.076	0.073	0.074	0.071
		RPU	0.079	0.072	0.072	0.070
	5 Items, Spread	None	0.057	0.058	0.059	0.053
		Delta	0.071	0.073	0.074	0.067
		IRT	0.072	0.072	0.073	0.067
		RPU	0.071	0.073	0.074	0.073

Continued, next page.

Table 4.9., cont'd.:

Table 4.9. Proportion of Examinees Under-classified, by Condition

			Negatively Skewed (Shift) Distribution *Degree of Aberrancy (a-shift/b-shift)*			
Administration 1 Condition	Condition Description	Detection Method	0.3 / 0.5	0.7 / 0.5	0.3 / 0.8	0.7 / 0.8
1	1 Item, Hard	None	0.064	0.064	0.062	0.063
		Delta	0.065	0.065	0.065	0.065
		IRT	0.065	0.065	0.065	0.065
		RPU	0.087	0.086	0.085	0.084
	3 Items, Spread	None	0.058	0.058	0.053	0.053
		Delta	0.059	0.059	0.059	0.059
		IRT	0.059	0.059	0.059	0.058
		RPU	0.085	0.084	0.084	0.083
	3 Items, Moderate	None	0.055	0.043	0.051	0.042
		Delta	0.056	0.055	0.057	0.054
		IRT	0.057	0.057	0.057	0.055
		RPU	0.079	0.076	0.080	0.081
	5 Items, Spread	None	0.051	0.045	0.043	0.041
		Delta	0.053	0.057	0.051	0.052
		IRT	0.052	0.057	0.050	0.053
		RPU	0.080	0.077	0.081	0.084
2	1 Item, Hard	None	0.062	0.063	0.061	0.061
		Delta	0.064	0.064	0.064	0.064
		IRT	0.064	0.064	0.064	0.064
		RPU	0.087	0.086	0.085	0.084
	3 Items, Spread	None	0.057	0.057	0.052	0.052
		Delta	0.058	0.058	0.058	0.058
		IRT	0.058	0.058	0.061	0.058
		RPU	0.086	0.085	0.084	0.083
	3 Items, Moderate	None	0.054	0.042	0.050	0.041
		Delta	0.055	0.055	0.056	0.053
		IRT	0.056	0.056	0.058	0.055
		RPU	0.079	0.078	0.081	0.085
	5 Items, Spread	None	0.050	0.044	0.043	0.040
		Delta	0.053	0.057	0.050	0.052
		IRT	0.052	0.056	0.054	0.052
		RPU	0.080	0.078	0.081	0.086

Table 4.10. Proportion of Examinees Over-classified, by Condition

			Normal (No Shift) Distribution			
			Degree of Aberrancy (a-shift/b-shift)			
Administration 1 Condition	Condition Description	Detection Method	0.3 / 0.5	0.7 / 0.5	0.3 / 0.8	0.7 / 0.8
1	1 Item, Hard	None	0.132	0.153	0.134	0.153
		Delta	0.129	0.129	0.129	0.129
		IRT	0.128	0.129	0.128	0.129
		RPU	0.132	0.129	0.134	0.129
	3 Items, Spread	None	0.137	0.143	0.145	0.146
		Delta	0.134	0.124	0.129	0.126
		IRT	0.134	0.131	0.136	0.134
		RPU	0.137	0.123	0.133	0.125
	3 Items, Moderate	None	0.145	0.157	0.156	0.159
		Delta	0.125	0.135	0.135	0.138
		IRT	0.133	0.135	0.134	0.138
		RPU	0.123	0.135	0.135	0.138
	5 Items, Spread	None	0.158	0.145	0.144	0.154
		Delta	0.137	0.125	0.123	0.132
		IRT	0.137	0.131	0.130	0.137
		RPU	0.137	0.124	0.123	0.126
2	1 Item, Hard	None	0.130	0.151	0.132	0.151
		Delta	0.128	0.128	0.127	0.128
		IRT	0.127	0.127	0.127	0.127
		RPU	0.130	0.128	0.132	0.128
	3 Items, Spread	None	0.135	0.142	0.143	0.145
		Delta	0.133	0.122	0.127	0.125
		IRT	0.133	0.129	0.133	0.132
		RPU	0.135	0.122	0.131	0.124
	3 Items, Moderate	None	0.143	0.155	0.155	0.157
		Delta	0.123	0.134	0.133	0.136
		IRT	0.131	0.133	0.133	0.136
		RPU	0.122	0.134	0.133	0.136
	5 Items, Spread	None	0.156	0.144	0.143	0.153
		Delta	0.136	0.123	0.122	0.131
		IRT	0.135	0.129	0.128	0.136
		RPU	0.135	0.123	0.122	0.125

Continued, next page.

Table 4.10., cont'd.:

Table 4.10. Proportion of Examinees Over-classified, by Condition

			Negatively Skewed (Shift) Distribution			
			Degree of Aberrancy (a-shift/b-shift)			
Administration 1 Condition	Condition Description	Detection Method	0.3 / 0.5	0.7 / 0.5	0.3 / 0.8	0.7 / 0.8
1	1 Item, Hard	None	0.176	0.175	0.179	0.178
		Delta	0.173	0.172	0.172	0.172
		IRT	0.171	0.170	0.171	0.171
		RPU	0.140	0.143	0.143	0.146
	3 Items, Spread	None	0.183	0.182	0.192	0.191
		Delta	0.179	0.179	0.178	0.178
		IRT	0.178	0.177	0.178	0.180
		RPU	0.140	0.144	0.143	0.145
	3 Items, Moderate	None	0.189	0.207	0.199	0.209
		Delta	0.186	0.180	0.184	0.183
		IRT	0.184	0.179	0.183	0.182
		RPU	0.148	0.142	0.146	0.140
	5 Items, Spread	None	0.193	0.190	0.210	0.200
		Delta	0.187	0.170	0.193	0.181
		IRT	0.188	0.171	0.192	0.180
		RPU	0.146	0.133	0.144	0.126
2	1 Item, Hard	None	0.173	0.172	0.176	0.175
		Delta	0.170	0.170	0.170	0.169
		IRT	0.168	0.168	0.168	0.168
		RPU	0.136	0.139	0.140	0.142
	3 Items, Spread	None	0.180	0.179	0.190	0.189
		Delta	0.177	0.177	0.177	0.177
		IRT	0.175	0.175	0.172	0.177
		RPU	0.137	0.139	0.141	0.143
	3 Items, Moderate	None	0.186	0.204	0.197	0.206
		Delta	0.184	0.178	0.183	0.181
		IRT	0.181	0.177	0.179	0.180
		RPU	0.144	0.139	0.143	0.137
	5 Items, Spread	None	0.190	0.189	0.207	0.198
		Delta	0.184	0.168	0.191	0.179
		IRT	0.185	0.169	0.184	0.178
		RPU	0.144	0.131	0.143	0.124

Table 4.11. Growth Expectations for the Negatively Skewed (Shift) Distribution Condition for Administration 2

			Administration 1 Condition 0.8 Correlated Thetas		
Condition Description	Degree of Aberrancy *(a-shift/b-shift)*	Detection Method	% Growth	% Difference From Expected	Growth Expectation Status
No Drift	None	None	9.7	▨	▨
1 Item, Hard	0.3 / 0.5	None	11.2	1.5	Meets
		Delta	10.8	1.1	Meets
		IRT	10.6	0.9	Meets
		RPU	5.3	-4.4	Under
	0.7 / 0.5	None	11.1	1.4	Meets
		Delta	10.7	1.0	Meets
		IRT	10.5	0.8	Meets
		RPU	5.6	-4.1	Under
	0.3 / 0.8	None	11.7	2.0	Over
		Delta	10.8	1.1	Meets
		IRT	10.7	1.0	Meets
		RPU	5.8	-3.9	Under
	0.7 / 0.8	None	11.5	1.8	Over
		Delta	10.7	1.0	Meets
		IRT	10.6	0.9	Meets
		RPU	6.1	-3.6	Under
3 Items, Spread	0.3 / 0.5	None	12.5	2.8	Over
		Delta	12.1	2.4	Over
		IRT	11.9	2.2	Over
		RPU	5.4	-4.3	Under
	0.7 / 0.5	None	12.4	2.7	Over
		Delta	12.0	2.3	Over
		IRT	11.8	2.1	Over
		RPU	6.0	-3.7	Under
	0.3 / 0.8	None	14.0	4.3	Over
		Delta	11.9	2.2	Over
		IRT	11.8	2.1	Over
		RPU	5.8	-3.9	Under
	0.7 / 0.8	None	13.8	4.1	Over
		Delta	12.0	2.3	Over
		IRT	12.2	2.5	Over
		RPU	6.2	-3.5	Under

Continued, next page.

Table 4.11., cont'd.:

Table 4.11. Growth Expectations for the Negatively Skewed (Shift) Distribution Condition for Administration 2

				Administration 1 Condition 0.8 Correlated Thetas	
Condition Description	Degree of Aberrancy *(a-shift/b-shift)*	Detection Method	% Growth	% Difference From Expected	Growth Expectation
No Drift	None	None	9.7	▨	▨
3 Items, Moderate	0.3 / 0.5	None	13.4	3.7	Over
		Delta	13.0	3.3	Over
		IRT	12.7	3.0	Over
		RPU	6.9	-2.8	Under
	0.7 / 0.5	None	16.4	6.7	Over
		Delta	12.5	2.8	Over
		IRT	12.3	2.6	Over
		RPU	6.6	-3.1	Under
	0.3 / 0.8	None	14.9	5.2	Over
		Delta	12.8	3.1	Over
		IRT	12.6	2.9	Over
		RPU	6.6	-3.1	Under
	0.7 / 0.8	None	16.7	7.0	Over
		Delta	12.9	3.2	Over
		IRT	12.7	3.0	Over
		RPU	5.9	-3.8	Under
5 Items, Spread	0.3 / 0.5	None	14.2	4.5	Over
		Delta	13.5	3.8	Over
		IRT	13.6	3.9	Over
		RPU	6.6	-3.1	Under
	0.7 / 0.5	None	14.6	4.9	Over
		Delta	11.3	1.6	Meets
		IRT	11.5	1.8	Over
		RPU	5.6	-4.1	Under
	0.3 / 0.8	None	16.7	7.0	Over
		Delta	14.2	4.5	Over
		IRT	14.2	4.5	Over
		RPU	6.3	-3.4	Under
	0.7 / 0.8	None	16.0	6.3	Over
		Delta	12.9	3.2	Over
		IRT	12.7	3.0	Over
		RPU	4.1	-5.6	Under

Continued, next page.

Table 4.11., cont'd.:

Table 4.11. Growth Expectations for the Negatively Skewed (Shift) Distribution Condition for Administration 2

			Administration 1 Condition 0.6 Correlated Thetas		
Condition Description	Degree of Aberrancy *(a-shift/b-shift)*	Detection Method	% Growth	% Difference From Expected	Growth Expectation Status
No Drift	None	None	9.5	▨	▨
1 Item, Hard	0.3 / 0.5	None	11.1	1.6	Meets
		Delta	10.7	1.2	Meets
		IRT	10.4	0.9	Meets
		RPU	4.9	-4.6	Under
	0.7 / 0.5	None	10.9	1.4	Meets
		Delta	10.6	1.1	Meets
		IRT	10.4	0.9	Meets
		RPU	5.2	-4.3	Under
	0.3 / 0.8	None	11.5	2.0	Over
		Delta	10.7	1.2	Meets
		IRT	10.4	0.9	Meets
		RPU	5.4	-4.1	Under
	0.7 / 0.8	None	11.4	1.9	Over
		Delta	10.5	1.0	Meets
		IRT	10.4	0.9	Meets
		RPU	5.7	-3.8	Under
3 Items, Spread	0.3 / 0.5	None	12.3	2.8	Over
		Delta	11.9	2.4	Over
		IRT	11.7	2.2	Over
		RPU	5.2	-4.3	Under
	0.7 / 0.5	None	12.2	2.7	Over
		Delta	11.9	2.4	Over
		IRT	11.8	2.3	Over
		RPU	5.4	-4.1	Under
	0.3 / 0.8	None	13.8	4.3	Over
		Delta	11.9	2.4	Over
		IRT	11.1	1.6	Meets
		RPU	5.7	-3.8	Under
	0.7 / 0.8	None	13.7	4.2	Over
		Delta	11.9	2.4	Over
		IRT	12.0	2.5	Over
		RPU	5.9	-3.6	Under

Table 4.11. Growth Expectations for the Negatively Skewed (Shift) Distribution Condition for Administration 2

			Administration 1 Condition 0.6 Correlated Thetas		
Condition Description	Degree of Aberrancy *(a-shift/b-shift)*	Detection Method	% Growth	% Difference From Expected	Growth Expectation
No Drift	None	None	9.5	▨	▨
3 Items, Moderate	0.3 / 0.5	None	13.2	3.7	Over
		Delta	12.8	3.3	Over
		IRT	12.6	3.1	Over
		RPU	6.5	-3.0	Under
	0.7 / 0.5	None	16.2	6.7	Over
		Delta	12.3	2.8	Over
		IRT	12.1	2.6	Over
		RPU	6.1	-3.4	Under
	0.3 / 0.8	None	14.7	5.2	Over
		Delta	12.7	3.2	Over
		IRT	12.1	2.6	Over
		RPU	6.3	-3.2	Under
	0.7 / 0.8	None	16.5	7.0	Over
		Delta	12.7	3.2	Over
		IRT	12.5	3.0	Over
		RPU	5.2	-4.3	Under
5 Items, Spread	0.3 / 0.5	None	14.0	4.5	Over
		Delta	13.1	3.6	Over
		IRT	13.2	3.7	Over
		RPU	6.4	-3.1	Under
	0.7 / 0.5	None	14.4	4.9	Over
		Delta	11.2	1.7	Meets
		IRT	11.3	1.8	Over
		RPU	5.3	-4.2	Under
	0.3 / 0.8	None	16.5	7.0	Over
		Delta	14.1	4.6	Over
		IRT	13.1	3.6	Over
		RPU	6.2	-3.3	Under
	0.7 / 0.8	None	15.8	6.3	Over
		Delta	12.8	3.3	Over
		IRT	12.6	3.1	Over
		RPU	3.8	-5.7	Under

Table 4.12. Grade 4 Delta Analysis

Item	*p* (Year 1)	*p* (Year 2)	Delta (Year 1)	Delta (Year 2)	Distance from Line	SD from Line	Flag (> 3 SDs)	Flag (> 2 SDs)
1	0.72	0.75	10.67	10.30	0.19	0.70	0	0
2	0.68	0.64	11.13	11.57	-0.38	1.43	0	0
3	0.49	0.52	13.10	12.80	0.14	0.53	0	0
4	0.79	0.80	9.77	9.63	0.03	0.10	0	0
5	0.63	0.61	11.67	11.88	-0.22	0.83	0	0
6	0.77	0.78	10.04	9.91	0.02	0.08	0	0
7	0.68	0.70	11.13	10.90	0.09	0.33	0	0
8	0.75	0.77	10.30	10.04	0.11	0.41	0	0
9	0.85	0.86	8.85	8.68	0.05	0.19	0	0
10	0.66	0.67	11.35	11.24	0.00	0.02	0	0
11	0.81	0.84	9.49	9.02	0.26	0.96	0	0
12	0.79	0.83	9.77	9.18	0.34	1.29	0	0
13	0.56	0.55	12.40	12.50	-0.14	0.54	0	0
14	0.87	0.88	8.49	8.30	0.06	0.24	0	0
15	0.72	0.76	10.67	10.17	0.28	1.03	0	0
16	0.53	0.51	12.70	12.90	-0.21	0.80	0	0
17	0.75	0.77	10.30	10.04	0.11	0.41	0	0
18	0.65	0.63	11.46	11.67	-0.22	0.84	0	0
19	0.63	0.67	11.67	11.24	0.23	0.87	0	0
20	0.81	0.78	9.49	9.91	-0.37	1.39	0	0
21	0.72	0.70	10.67	10.90	-0.24	0.89	0	0
22	0.51	0.63	12.90	11.67	0.80	2.98	0	1
23	0.84	0.82	9.02	9.34	-0.30	1.11	0	0
24	0.74	0.75	10.43	10.30	0.02	0.06	0	0
25	0.93	0.93	7.10	7.10	-0.07	0.28	0	0
26	0.60	0.52	11.99	12.80	-0.65	2.42	0	1
27	0.65	0.64	11.46	11.57	-0.15	0.56	0	0
28	0.50	0.56	13.00	12.40	0.35	1.33	0	0
29	0.72	0.75	10.67	10.30	0.19	0.70	0	0
30	0.92	0.92	7.38	7.38	-0.07	0.28	0	0
31	0.71	0.73	10.79	10.55	0.10	0.36	0	0
32	0.85	0.83	8.85	9.18	-0.31	1.15	0	0
33	0.58	0.59	12.19	12.09	0.00	0.00	0	0
34	0.54	0.55	12.60	12.50	0.00	0.00	0	0
35	0.56	0.61	12.40	11.88	0.29	1.09	0	0
36	0.48	0.46	13.20	13.40	-0.21	0.80	0	0
37	0.43	0.41	13.76	13.94	-0.20	0.74	0	0
38	0.41	0.47	13.88	13.28	0.36	1.34	0	0
39	0.49	0.46	13.13	13.38	-0.25	0.93	0	0

Table 4.13. Grade 8 Delta Analysis

Item	*p* (Year 1)	*p* (Year 2)	Delta (Year 1)	Delta (Year 2)	Distance from Line	SD from Line	Flag (> 3 SDs)	Flag (> 2 SDs)
1	0.70	0.68	10.90	10.30	0.33	1.81	0	0
2	0.70	0.70	10.90	10.90	-0.09	0.50	0	0
3	0.46	0.46	13.40	13.40	-0.03	0.16	0	0
4	0.70	0.70	10.90	10.90	-0.09	0.50	0	0
5	0.47	0.45	13.30	13.50	-0.17	0.95	0	0
6	0.65	0.71	11.46	10.79	0.39	2.17	0	1
7	0.59	0.58	12.09	12.19	-0.13	0.73	0	0
8	0.57	0.59	12.29	12.09	0.09	0.48	0	0
9	0.69	0.68	11.02	11.13	-0.17	0.92	0	0
10	0.51	0.51	12.90	12.90	-0.04	0.23	0	0
11	0.39	0.40	14.12	14.01	0.06	0.34	0	0
12	0.81	0.81	9.49	9.49	-0.13	0.70	0	0
13	0.68	0.70	11.13	10.90	0.07	0.40	0	0
14	0.64	0.66	11.57	11.35	0.08	0.42	0	0
15	0.73	0.74	10.55	10.43	-0.01	0.08	0	0
16	0.62	0.65	11.78	11.46	0.15	0.85	0	0
17	0.60	0.59	11.99	12.09	-0.14	0.75	0	0
18	0.48	0.52	13.20	12.80	0.25	1.36	0	0
19	0.53	0.55	12.70	12.50	0.09	0.52	0	0
20	0.48	0.48	13.20	13.20	-0.03	0.18	0	0
21	0.40	0.42	14.01	13.81	0.13	0.72	0	0
22	0.83	0.84	9.18	9.02	-0.02	0.12	0	0
23	0.69	0.72	11.02	10.67	0.15	0.86	0	0
24	0.82	0.82	9.34	9.34	-0.13	0.72	0	0
25	0.50	0.52	13.00	12.80	0.10	0.56	0	0
26	0.66	0.64	11.35	11.57	-0.23	1.27	0	0
27	0.83	0.84	9.18	9.02	-0.02	0.12	0	0
28	0.67	0.62	11.24	11.78	-0.46	2.53	0	1
29	0.49	0.48	13.10	13.20	-0.11	0.59	0	0
30	0.80	0.81	9.63	9.49	-0.02	0.12	0	0
31	0.67	0.69	11.24	11.02	0.07	0.41	0	0
32	0.80	0.83	9.63	9.18	0.19	1.06	0	0
33	0.48	0.49	13.20	13.10	0.04	0.20	0	0
34	0.64	0.66	11.57	11.35	0.08	0.42	0	0
35	0.28	0.28	15.33	15.33	0.02	0.11	0	0
36	0.58	0.56	12.19	12.40	-0.20	1.11	0	0
37	0.45	0.45	13.55	13.55	-0.02	0.14	0	0
38	0.42	0.37	13.78	14.35	-0.42	2.31	0	1
39	0.37	0.42	14.35	13.83	0.36	1.98	0	0

Table 4.14. Grade 11 Delta Analysis

Item	*p* (Year 1)	*p* (Year 2)	Delta (Year 1)	Delta (Year 2)	Distance from Line	SD from Line	Flag (> 3 SDs)	Flag (> 2 SDs)
1	0.41	0.46	13.91	13.40	0.27	1.39	0	0
2	0.53	0.50	12.70	13.00	-0.28	1.45	0	0
3	0.53	0.50	12.70	13.00	-0.28	1.45	0	0
4	0.79	0.81	9.77	9.49	0.22	1.14	0	0
5	0.58	0.61	12.19	11.88	0.17	0.89	0	0
6	0.70	0.69	10.90	11.02	-0.10	0.51	0	0
7	0.54	0.55	12.60	12.50	0.01	0.06	0	0
8	0.36	0.39	14.43	14.12	0.12	0.60	0	0
9	0.65	0.66	11.46	11.35	0.05	0.24	0	0
10	0.70	0.71	10.90	10.79	0.07	0.35	0	0
11	0.40	0.39	14.01	14.12	-0.17	0.90	0	0
12	0.70	0.65	10.90	11.46	-0.42	2.15	0	1
13	0.57	0.58	12.29	12.19	0.02	0.10	0	0
14	0.51	0.52	12.90	12.80	0.00	0.01	0	0
15	0.37	0.37	14.33	14.33	-0.11	0.56	0	0
16	0.51	0.55	12.90	12.50	0.22	1.14	0	0
17	0.60	0.58	11.99	12.19	-0.19	1.00	0	0
18	0.79	0.77	9.77	10.04	-0.18	0.93	0	0
19	0.73	0.72	10.55	10.67	-0.09	0.48	0	0
20	0.46	0.48	13.40	13.20	0.06	0.32	0	0
21	0.36	0.35	14.43	14.54	-0.19	0.97	0	0
22	0.31	0.36	14.98	14.43	0.27	1.39	0	0
23	0.30	0.30	15.10	15.10	-0.13	0.67	0	0
24	0.73	0.73	10.55	10.55	-0.01	0.03	0	0
25	0.48	0.50	13.20	13.00	0.07	0.34	0	0
26	0.61	0.58	11.88	12.19	-0.27	1.37	0	0
27	0.56	0.57	12.40	12.29	0.02	0.09	0	0
28	0.45	0.46	13.50	13.40	-0.01	0.07	0	0
29	0.65	0.64	11.46	11.57	-0.11	0.56	0	0
30	0.59	0.62	12.09	11.78	0.18	0.91	0	0
31	0.59	0.61	12.09	11.88	0.10	0.52	0	0
32	0.31	0.28	14.98	15.33	-0.38	1.94	0	0
33	0.66	0.70	11.35	10.90	0.29	1.52	0	0
34	0.83	0.83	9.18	9.18	0.03	0.16	0	0
35	0.65	0.65	11.46	11.46	-0.03	0.16	0	0
36	0.65	0.67	11.46	11.24	0.13	0.65	0	0
37	0.29	0.35	15.21	14.60	0.31	1.62	0	0
38	0.24	0.25	15.83	15.67	-0.03	0.18	0	0
39	0.39	0.45	14.17	13.50	0.38	1.94	0	0

Table 4.15. Grade 4 *b*-Parameter Drift Analysis

Item	Item Type	*b* (Year 2)	*b* (Year 1)	Distance from Line	SD from Line	Flag (> 2 SDs)
1	Multiple Choice	-0.72	-0.51	0.01	0.05	0
2	Multiple Choice	-0.99	-0.37	-0.28	1.04	0
3	Multiple Choice	0.30	0.34	0.13	0.50	0
4	Multiple Choice	-1.65	-1.66	0.16	0.60	0
5	Multiple Choice	-0.45	0.00	-0.16	0.60	0
6	Multiple Choice	-1.17	-1.05	0.07	0.26	0
7	Multiple Choice	-0.69	-0.60	0.09	0.35	0
8	Multiple Choice	-0.96	-0.47	-0.19	0.70	0
9	Multiple Choice	-1.46	-1.38	0.10	0.39	0
10	Multiple Choice	-0.29	-0.20	0.10	0.36	0
11	Multiple Choice	-1.54	-1.46	0.11	0.41	0
12	Multiple Choice	-0.56	-1.54	0.86	3.20	1
13	Multiple Choice	-0.20	0.21	-0.13	0.48	0
14	Multiple Choice	-2.39	-2.33	0.12	0.44	0
15	Multiple Choice	-0.74	-0.64	0.10	0.36	0
16	Multiple Choice	0.16	0.56	-0.12	0.44	0
17	Multiple Choice	-1.30	-0.96	-0.08	0.29	0
18	Multiple Choice	-0.55	-0.24	-0.06	0.24	0
19	Multiple Choice	-0.58	-0.52	0.12	0.45	0
20	Multiple Choice	-2.54	-1.70	-0.43	1.62	0
21	Multiple Choice	-1.15	-0.75	-0.12	0.46	0
22	Multiple Choice	0.35	0.24	0.23	0.88	0
23	Multiple Choice	-3.27	-2.50	-0.39	1.45	0
24	Multiple Choice	-0.87	-0.48	-0.12	0.43	0
25	Multiple Choice	-3.34	-3.17	0.04	0.15	0
26	Multiple Choice	0.30	1.60	-0.76	2.84	1
27	Multiple Choice	-0.20	0.46	-0.31	1.15	0
28	Multiple Choice	0.29	0.21	0.21	0.79	0
29	Multiple Choice	-0.84	-0.08	-0.38	1.41	0
30	Multiple Choice	-2.41	-2.32	0.10	0.37	0
31	Multiple Choice	-0.71	-0.74	0.19	0.70	0
32	Multiple Choice	-1.79	-1.39	-0.12	0.46	0
33	Multiple Choice	0.18	0.44	-0.03	0.10	0
34	Multiple Choice	-0.12	0.11	0.00	0.01	0
35	Multiple Choice	0.09	0.08	0.17	0.62	0
36	Multiple Choice	0.51	0.87	-0.10	0.36	0
37	Open Response	0.63	0.37	0.35	1.29	0
38	Open Response	0.31	0.40	0.10	0.36	0
39	Open Response	0.12	-0.24	0.42	1.55	0

Table 4.16. Grade 8 *b*-Parameter Drift Analysis

Item	Item Type	*b* (Year 2)	*b* (Year 1)	Distance from Line	SD from Line	Flag (> 2 SDs)
1	Multiple Choice	-0.505	-0.776	0.227	1.549	0
2	Multiple Choice	-0.650	-0.508	-0.063	0.428	0
3	Multiple Choice	0.474	0.596	-0.013	0.086	0
4	Multiple Choice	-0.714	-0.704	0.026	0.178	0
5	Multiple Choice	0.995	0.951	0.120	0.820	0
6	Multiple Choice	-1.006	-0.669	-0.209	1.428	0
7	Multiple Choice	0.081	0.242	-0.052	0.354	0
8	Multiple Choice	0.237	0.519	-0.130	0.890	0
9	Multiple Choice	-0.622	-0.555	-0.010	0.067	0
10	Multiple Choice	0.447	0.245	0.210	1.437	0
11	Multiple Choice	1.125	1.426	-0.115	0.782	0
12	Multiple Choice	-1.469	-1.552	0.066	0.448	0
13	Multiple Choice	-0.848	-0.758	-0.034	0.232	0
14	Multiple Choice	-0.461	-0.433	0.022	0.150	0
15	Multiple Choice	-0.755	-0.869	0.110	0.751	0
16	Multiple Choice	-0.371	-0.077	-0.159	1.085	0
17	Multiple Choice	0.065	-0.306	0.315	2.148	1
18	Multiple Choice	0.610	1.079	-0.247	1.688	0
19	Multiple Choice	0.317	0.472	-0.041	0.277	0
20	Multiple Choice	0.546	0.626	0.019	0.132	0
21	Multiple Choice	0.803	0.882	0.028	0.194	0
22	Multiple Choice	-0.913	-1.429	0.383	2.613	1
23	Multiple Choice	-1.050	-0.832	-0.129	0.879	0
24	Multiple Choice	-1.672	-1.737	0.047	0.318	0
25	Multiple Choice	0.499	0.717	-0.078	0.529	0
26	Multiple Choice	0.424	0.221	0.210	1.436	0
27	Multiple Choice	-1.920	-1.809	-0.082	0.562	0
28	Multiple Choice	0.243	0.210	0.087	0.594	0
29	Multiple Choice	0.691	0.752	0.037	0.253	0
30	Multiple Choice	-1.575	-1.214	-0.245	1.672	0
31	Multiple Choice	-0.488	-0.410	-0.013	0.090	0
32	Multiple Choice	-1.711	-1.511	-0.137	0.939	0
33	Multiple Choice	0.278	0.657	-0.196	1.339	0
34	Multiple Choice	-0.281	-0.100	-0.078	0.531	0
35	Multiple Choice	1.772	1.889	0.034	0.230	0
36	Multiple Choice	-0.134	-0.303	0.169	1.154	0
37	Open Response	0.133	0.274	-0.036	0.248	0
38	Open Response	0.473	0.430	0.102	0.697	0
39	Open Response	0.213	0.516	-0.146	0.994	0

Table 4.17. Grade 11 *b*-Parameter Drift Analysis

Item	Item Type	*b* (Year 2)	*b* (Year 1)	Distance from Line	SD from Line	Flag (> 2 SDs)
1	Multiple Choice	0.41	0.47	-0.04	0.13	0
2	Multiple Choice	-1.35	-1.37	-0.17	0.60	0
3	Multiple Choice	0.57	-0.38	0.64	2.22	1
4	Multiple Choice	-0.97	-1.00	-0.13	0.44	0
5	Multiple Choice	1.14	1.82	-0.37	1.28	0
6	Multiple Choice	0.50	0.15	0.24	0.84	0
7	Multiple Choice	1.37	1.33	0.12	0.42	0
8	Multiple Choice	0.02	-0.03	-0.01	0.05	0
9	Multiple Choice	1.47	1.75	-0.07	0.24	0
10	Multiple Choice	-0.24	0.29	-0.41	1.43	0
11	Multiple Choice	0.63	0.73	-0.04	0.16	0
12	Multiple Choice	-0.12	-0.15	-0.04	0.14	0
13	Multiple Choice	-0.42	-0.36	-0.13	0.45	0
14	Multiple Choice	1.05	0.02	0.73	2.54	1
15	Multiple Choice	1.81	1.53	0.32	1.12	0
16	Multiple Choice	-1.01	-1.12	-0.08	0.28	0
17	Multiple Choice	1.23	1.23	0.09	0.30	0
18	Multiple Choice	-0.70	-0.88	0.00	0.01	0
19	Multiple Choice	1.35	1.29	0.14	0.49	0
20	Multiple Choice	0.63	0.78	-0.07	0.25	0
21	Multiple Choice	0.19	-0.02	0.11	0.39	0
22	Multiple Choice	-1.06	-1.26	-0.02	0.07	0
23	Multiple Choice	1.06	1.08	0.06	0.19	0
24	Multiple Choice	-0.29	-0.31	-0.06	0.22	0
25	Multiple Choice	-0.50	-0.46	-0.13	0.44	0
26	Multiple Choice	0.65	0.77	-0.05	0.19	0
27	Multiple Choice	0.98	1.18	-0.07	0.24	0
28	Multiple Choice	0.85	0.73	0.12	0.43	0
29	Multiple Choice	-0.59	-0.95	0.13	0.46	0
30	Multiple Choice	0.17	0.10	0.02	0.07	0
31	Multiple Choice	0.12	-0.02	0.06	0.21	0
32	Multiple Choice	0.96	0.85	0.13	0.44	0
33	Multiple Choice	0.57	0.61	-0.01	0.04	0
34	Multiple Choice	0.01	-0.26	0.13	0.46	0
35	Multiple Choice	0.86	0.70	0.15	0.52	0
36	Multiple Choice	-0.46	-0.70	0.06	0.21	0
37	Open Response	0.44	2.28	-1.19	4.15	1
38	Open Response	1.14	1.18	0.05	0.18	0
39	Open Response	0.32	0.62	-0.21	0.71	0

Table 4.18. Grade 4 RPU Analysis

Item	Item Type	Score Point	RPU 1 (>0.07)	RPU 2 (>0.07)	RPU 3 (>0.07)	RPU 4 (>0.07)	RPU 5 (>0.07)	RPU 6 (>0.07)	RPU 7 (>0.125)	Flag (> 4 Flags)
1	Multiple Choice	1	-0.03	0.03	0.03	0.00	0.00	0.00	0.023	0
2	Multiple Choice	1	-0.03	0.03	0.04	0.00	0.00	0.00	0.029	0
3	Multiple Choice	1	-0.05	0.05	0.06	0.00	0.00	0.01	0.042	0
4	Multiple Choice	1	-0.04	0.04	0.05	0.00	0.00	0.01	0.060	0
5	Multiple Choice	1	-0.01	0.02	0.03	0.00	0.00	0.00	0.001	0
6	Multiple Choice	1	-0.02	0.02	0.03	0.00	0.00	0.01	0.053	0
7	Multiple Choice	1	0.07	0.07	0.09	0.00	0.00	0.01	0.088	0
8	Multiple Choice	1	-0.02	0.02	0.02	0.00	0.00	0.00	0.027	0
9	Multiple Choice	1	0.00	0.01	0.01	0.00	0.00	0.00	0.002	0
10	Multiple Choice	1	-0.04	0.04	0.04	0.00	0.00	0.01	0.061	0
11	Multiple Choice	1	-0.02	0.02	0.02	0.00	0.00	0.00	0.030	0
12	Multiple Choice	1	-0.01	0.01	0.02	0.00	0.00	0.00	0.013	0
13	Multiple Choice	1	-0.03	0.03	0.03	0.00	0.00	0.00	0.035	0
14	Multiple Choice	1	-0.05	0.05	0.06	0.00	0.00	0.01	0.082	0
15	Multiple Choice	1	-0.03	0.04	0.04	0.00	0.00	0.01	0.056	0
16	Multiple Choice	1	-0.04	0.04	0.04	0.00	0.00	0.01	0.051	0
17	Multiple Choice	1	0.00	0.02	0.03	0.00	0.00	0.00	0.029	0
18	Multiple Choice	1	-0.02	0.02	0.02	0.00	0.00	0.00	0.023	0
19	Multiple Choice	1	-0.05	0.05	0.06	0.00	0.00	0.01	0.046	0
20	Multiple Choice	1	-0.08	0.08	0.08	0.00	0.00	0.01	0.091	0
21	Multiple Choice	1	0.00	0.01	0.01	0.00	0.00	0.00	0.004	0
22	Multiple Choice	1	-0.02	0.02	0.02	0.00	0.00	0.00	0.014	0
23	Multiple Choice	1	-0.02	0.02	0.02	0.00	0.00	0.00	0.019	0
24	Multiple Choice	1	-0.05	0.05	0.06	0.00	0.00	0.01	0.052	0
25	Multiple Choice	1	0.02	0.04	0.05	0.00	0.00	0.01	0.029	0
26	Multiple Choice	1	0.00	0.01	0.01	0.00	0.00	0.00	0.012	0
27	Multiple Choice	1	0.00	0.03	0.04	0.00	0.00	0.00	0.010	0
28	Multiple Choice	1	0.00	0.02	0.03	0.00	0.00	0.00	0.022	0
29	Multiple Choice	1	-0.04	0.04	0.04	0.00	0.00	0.01	0.046	0
30	Multiple Choice	1	0.01	0.01	0.01	0.00	0.00	0.00	0.007	0
31	Multiple Choice	1	0.00	0.02	0.03	0.00	0.00	0.00	0.007	0
32	Multiple Choice	1	-0.01	0.02	0.02	0.00	0.00	0.00	0.021	0
33	Multiple Choice	1	-0.07	0.07	0.07	0.00	0.00	0.01	0.075	0
34	Multiple Choice	1	-0.01	0.01	0.01	0.00	0.00	0.00	0.013	0
35	Multiple Choice	1	-0.03	0.03	0.04	0.00	0.00	0.01	0.069	0
36	Multiple Choice	1	-0.01	0.01	0.01	0.00	0.00	0.00	0.010	0
37	Open Response	1	-0.05	0.05	0.06	0.00	0.00	0.01	0.045	0
37	Open Response	2	-0.07	0.07	0.08	0.00	0.00	0.01	0.120	0
37	Open Response	3	-0.05	0.05	0.06	0.00	0.00	0.01	0.076	0
37	Open Response	4	-0.02	0.02	0.02	0.00	0.00	0.00	0.014	0
38	Open Response	1	0.00	0.00	0.01	0.00	0.00	0.00	0.002	0
38	Open Response	2	0.03	0.03	0.03	0.00	0.00	0.01	0.048	0
38	Open Response	3	0.02	0.02	0.03	0.00	0.00	0.00	0.026	0
38	Open Response	4	0.01	0.01	0.01	0.00	0.00	0.00	0.003	0
39	Open Response	1	-0.02	0.02	0.03	0.00	0.00	0.00	0.017	0
39	Open Response	2	-0.11	0.11	0.13	0.00	0.00	0.02	0.180	1
39	Open Response	3	-0.06	0.07	0.09	0.00	0.00	0.01	0.076	0
39	Open Response	4	-0.04	0.04	0.07	0.00	0.00	0.01	0.007	0

Table 4.19. Grade 8 RPU Analysis

		Score	RPU 1	RPU 2	RPU 3	RPU 4	RPU 5	RPU 6	RPU 7	Flag
Item	Item Type	Point	(>0.07)	(>0.07)	(>0.07)	(>0.07)	(>0.07)	(>0.07)	(>0.125)	(> 4 Flags)
1	Multiple Choice	1	0.02	0.03	0.03	0.00	0.00	0.00	0.040	0
2	Multiple Choice	1	0.01	0.01	0.01	0.00	0.00	0.00	0.017	0
3	Multiple Choice	1	0.00	0.01	0.01	0.00	0.00	0.00	0.013	0
4	Multiple Choice	1	0.03	0.03	0.03	0.00	0.00	0.00	0.047	0
5	Multiple Choice	1	0.05	0.05	0.06	0.00	0.00	0.01	0.044	0
6	Multiple Choice	1	0.04	0.04	0.04	0.00	0.00	0.01	0.054	0
7	Multiple Choice	1	0.03	0.03	0.04	0.00	0.00	0.00	0.021	0
8	Multiple Choice	1	0.02	0.02	0.02	0.00	0.00	0.00	0.022	0
9	Multiple Choice	1	0.06	0.06	0.06	0.00	0.00	0.01	0.079	0
10	Multiple Choice	1	0.04	0.04	0.05	0.00	0.00	0.01	0.038	0
11	Multiple Choice	1	0.04	0.04	0.04	0.00	0.00	0.01	0.043	0
12	Multiple Choice	1	0.02	0.03	0.03	0.00	0.00	0.00	0.013	0
13	Multiple Choice	1	0.01	0.01	0.01	0.00	0.00	0.00	0.004	0
14	Multiple Choice	1	0.05	0.05	0.05	0.00	0.00	0.01	0.052	0
15	Multiple Choice	1	0.03	0.03	0.03	0.00	0.00	0.01	0.048	0
16	Multiple Choice	1	0.01	0.01	0.01	0.00	0.00	0.00	0.016	0
17	Multiple Choice	1	0.02	0.02	0.02	0.00	0.00	0.00	0.010	0
18	Multiple Choice	1	0.02	0.02	0.02	0.00	0.00	0.00	0.003	0
19	Multiple Choice	1	0.02	0.03	0.03	0.00	0.00	0.01	0.057	0
20	Multiple Choice	1	0.04	0.04	0.05	0.00	0.00	0.01	0.021	0
21	Multiple Choice	1	0.01	0.02	0.02	0.00	0.00	0.00	0.010	0
22	Multiple Choice	1	-0.03	0.03	0.04	0.00	0.00	0.01	0.036	0
23	Multiple Choice	1	0.06	0.06	0.07	0.00	0.00	0.01	0.064	0
24	Multiple Choice	1	0.01	0.01	0.01	0.00	0.00	0.00	0.007	0
25	Multiple Choice	1	0.01	0.01	0.01	0.00	0.00	0.00	0.006	0
26	Multiple Choice	1	0.02	0.02	0.03	0.00	0.00	0.00	0.013	0
27	Multiple Choice	1	0.00	0.01	0.01	0.00	0.00	0.00	0.008	0
28	Multiple Choice	1	0.03	0.03	0.04	0.00	0.00	0.00	0.027	0
29	Multiple Choice	1	0.04	0.04	0.05	0.00	0.00	0.01	0.093	0
30	Multiple Choice	1	0.03	0.03	0.04	0.00	0.00	0.01	0.043	0
31	Multiple Choice	1	0.00	0.03	0.04	0.00	0.00	0.00	0.025	0
32	Multiple Choice	1	0.01	0.02	0.02	0.00	0.00	0.00	0.037	0
33	Multiple Choice	1	0.04	0.04	0.04	0.00	0.00	0.01	0.030	0
34	Multiple Choice	1	0.02	0.02	0.03	0.00	0.00	0.00	0.013	0
35	Multiple Choice	1	0.04	0.04	0.05	0.00	0.00	0.01	0.039	0
36	Multiple Choice	1	0.02	0.02	0.03	0.00	0.00	0.00	0.006	0
37	Open Response	1	-0.05	0.05	0.06	0.00	0.00	0.01	0.027	0
37	Open Response	2	0.01	0.01	0.02	0.00	0.00	0.00	0.015	0
37	Open Response	3	0.06	0.06	0.07	0.00	0.00	0.01	0.091	0
37	Open Response	4	0.06	0.06	0.09	0.00	0.00	0.01	0.022	0
38	Open Response	1	-0.01	0.01	0.01	0.00	0.00	0.00	0.007	0
38	Open Response	2	-0.08	0.08	0.10	0.00	0.00	0.02	0.170	1
38	Open Response	3	-0.04	0.04	0.05	0.00	0.00	0.01	0.053	0
38	Open Response	4	0.05	0.05	0.09	0.00	0.00	0.01	0.016	0
39	Open Response	1	0.06	0.06	0.07	0.00	0.00	0.01	0.088	0
39	Open Response	2	0.06	0.06	0.08	0.00	0.00	0.02	0.152	0
39	Open Response	3	0.04	0.04	0.06	0.00	0.00	0.01	0.077	0
39	Open Response	4	0.05	0.05	0.07	0.00	0.00	0.01	0.047	0

Table 4.20. Grade 11 RPU Analysis

Item	Item Type	Score Point	RPU 1 (>0.07)	RPU 2 (>0.07)	RPU 3 (>0.07)	RPU 4 (>0.07)	RPU 5 (>0.07)	RPU 6 (>0.07)	RPU 7 (>0.125)	Flag (> 4 Flags)
1	Multiple Choice	1	0.05	0.05	0.05	0.00	0.00	0.01	0.040	0
2	Multiple Choice	1	-0.03	0.04	0.05	0.00	0.00	0.01	0.032	0
3	Multiple Choice	1	-0.01	0.03	0.04	0.00	0.00	0.01	0.057	0
4	Multiple Choice	1	0.02	0.02	0.02	0.00	0.00	0.00	0.012	0
5	Multiple Choice	1	0.03	0.04	0.06	0.00	0.00	0.01	0.020	0
6	Multiple Choice	1	0.00	0.04	0.04	0.00	0.00	0.01	0.029	0
7	Multiple Choice	1	0.00	0.00	0.00	0.00	0.00	0.00	0.006	0
8	Multiple Choice	1	0.02	0.03	0.04	0.00	0.00	0.00	0.030	0
9	Multiple Choice	1	0.02	0.02	0.03	0.00	0.00	0.00	0.010	0
10	Multiple Choice	1	0.02	0.02	0.03	0.00	0.00	0.00	0.001	0
11	Multiple Choice	1	0.00	0.01	0.02	0.00	0.00	0.00	0.026	0
12	Multiple Choice	1	-0.05	0.05	0.06	0.00	0.00	0.01	0.049	0
13	Multiple Choice	1	0.00	0.01	0.01	0.00	0.00	0.00	0.006	0
14	Multiple Choice	1	0.01	0.01	0.02	0.00	0.00	0.00	0.021	0
15	Multiple Choice	1	0.01	0.01	0.02	0.00	0.00	0.00	0.003	0
16	Multiple Choice	1	0.02	0.02	0.02	0.00	0.00	0.00	0.034	0
17	Multiple Choice	1	-0.02	0.02	0.03	0.00	0.00	0.00	0.038	0
18	Multiple Choice	1	0.00	0.02	0.03	0.00	0.00	0.00	0.016	0
19	Multiple Choice	1	0.00	0.01	0.02	0.00	0.00	0.00	0.016	0
20	Multiple Choice	1	0.03	0.03	0.03	0.00	0.00	0.00	0.010	0
21	Multiple Choice	1	-0.02	0.02	0.02	0.00	0.00	0.00	0.007	0
22	Multiple Choice	1	0.04	0.04	0.04	0.00	0.00	0.01	0.044	0
23	Multiple Choice	1	-0.01	0.01	0.01	0.00	0.00	0.00	0.003	0
24	Multiple Choice	1	0.01	0.02	0.03	0.00	0.00	0.00	0.009	0
25	Multiple Choice	1	0.01	0.01	0.02	0.00	0.00	0.00	0.024	0
26	Multiple Choice	1	-0.02	0.03	0.04	0.00	0.00	0.01	0.056	0
27	Multiple Choice	1	0.00	0.01	0.01	0.00	0.00	0.00	0.013	0
28	Multiple Choice	1	0.00	0.02	0.02	0.00	0.00	0.00	0.025	0
29	Multiple Choice	1	0.00	0.03	0.03	0.00	0.00	0.00	0.039	0
30	Multiple Choice	1	0.03	0.03	0.04	0.00	0.00	0.01	0.029	0
31	Multiple Choice	1	0.02	0.04	0.05	0.00	0.00	0.01	0.005	0
32	Multiple Choice	1	-0.02	0.04	0.05	0.00	0.00	0.01	0.041	0
33	Multiple Choice	1	0.04	0.04	0.06	0.00	0.00	0.01	0.019	0
34	Multiple Choice	1	0.00	0.00	0.01	0.00	0.00	0.00	0.000	0
35	Multiple Choice	1	0.00	0.01	0.01	0.00	0.00	0.00	0.004	0
36	Multiple Choice	1	0.02	0.02	0.02	0.00	0.00	0.00	0.012	0
37	Open Response	1	0.08	0.08	0.10	0.00	0.00	0.01	0.107	0
37	Open Response	2	0.11	0.11	0.12	0.00	0.00	0.02	0.179	1
37	Open Response	3	0.00	0.01	0.01	0.00	0.00	0.00	0.008	0
37	Open Response	4	0.01	0.01	0.01	0.00	0.00	0.00	0.001	0
38	Open Response	1	0.01	0.01	0.02	0.00	0.00	0.00	0.034	0
38	Open Response	2	0.00	0.01	0.01	0.00	0.00	0.00	0.006	0
38	Open Response	3	0.00	0.01	0.01	0.00	0.00	0.00	0.006	0
38	Open Response	4	0.02	0.02	0.03	0.00	0.00	0.00	0.004	0
39	Open Response	1	0.03	0.03	0.04	0.00	0.00	0.01	0.035	0
39	Open Response	2	0.07	0.07	0.09	0.00	0.00	0.02	0.152	1
39	Open Response	3	0.05	0.05	0.07	0.00	0.00	0.01	0.063	0
39	Open Response	4	0.03	0.03	0.06	0.00	0.00	0.00	0.008	0

Table 4.21. Scaling Constants for Each of the Detection Methods, by Grade

Grade	Detection Method	A	B
4	Delta (3 SDs)	0.991	-0.175
	Delta (2 SDs)	0.996	-0.182
	b-Parameter	0.992	-0.170
	RPU	1.011	-0.156
8	Delta (3 SDs)	1.030	0.138
	Delta (2 SDs)	1.030	0.135
	b-Parameter	1.020	0.139
	RPU	1.030	0.172
11	Delta (3 SDs)	0.966	0.068
	Delta (2 SDs)	0.961	0.077
	b-Parameter	0.968	0.045
	RPU	0.963	0.012

Table 4.22. Proportion of Examinees Classified into Each Proficiency Category & Differences between Detection Methods, Grade 4

	Detection Method	Proficiency Category				MEETS AYP	Δ_{Meets}	
		1	2	3	4		From Year 1	From Delta (3SD)
Year 1	▨	0.236	0.278	0.445	0.042	0.487	▨	▨
Year 2	Delta (3SD)	0.212	0.273	0.480	0.036	0.516	0.029	▨
	Delta (2SD)	0.215	0.271	0.477	0.036	0.513	0.026	-0.003
	b-Parameter	0.211	0.271	0.481	0.036	0.517	0.030	0.001
	RPU	0.210	0.265	0.482	0.042	0.524	0.037	0.008

Table 4.23. Proportion of Examinees Classified into Each Proficiency Category & Differences between Detection Methods, Grade 8

	Detection Method	Proficiency Category				MEETS AYP	Δ_{Meets}	
		1	2	3	4		From Year 1	From Delta (3SD)
Year 1		0.231	0.269	0.454	0.045	0.499		
Year 2	Delta (3SD)	0.240	0.265	0.454	0.041	0.495	-0.004	
	Delta (2SD)	0.228	0.259	0.467	0.046	0.513	0.014	0.018
	b-Parameter	0.226	0.259	0.468	0.046	0.514	0.015	0.019
	RPU	0.218	0.254	0.477	0.051	0.528	0.029	0.033

Table 4.24. Proportion of Examinees Classified into Each Proficiency Category & Differences between Detection Methods, Grade 11

	Detection Method	Proficiency Category				MEETS AYP	Δ_{Meets}	
		1	2	3	4		From Year 1	From Delta (3SD)
Year 1		0.216	0.260	0.482	0.043	0.525		
Year 2	Delta (3SD)	0.194	0.250	0.513	0.043	0.556	0.031	
	Delta (2SD)	0.190	0.250	0.517	0.043	0.560	0.035	0.004
	b-Parameter	0.200	0.253	0.507	0.040	0.547	0.022	-0.009
	RPU	0.207	0.258	0.500	0.035	0.535	0.010	-0.021

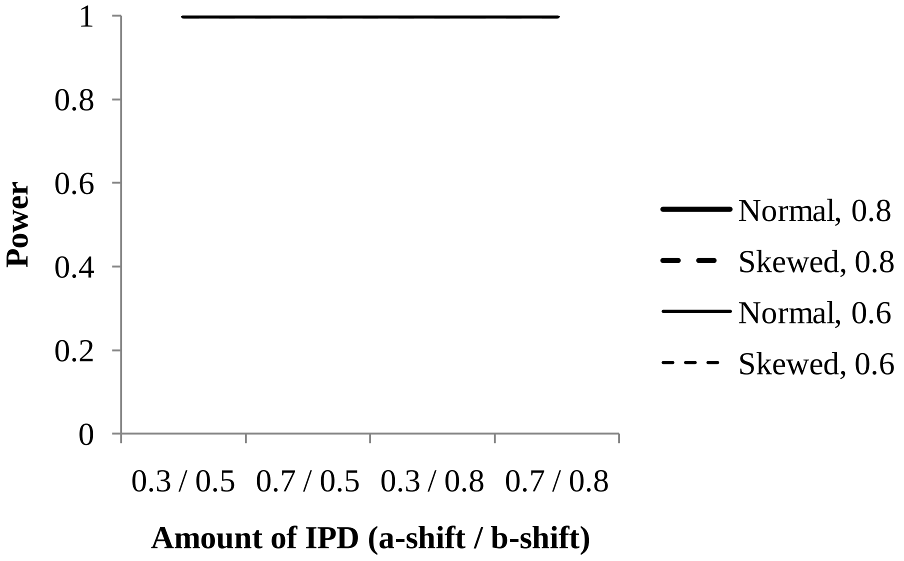


Figure 4.1. Power for Varying Levels of IPD for each Distribution Condition, 1 Hard Item, Delta Plot Method

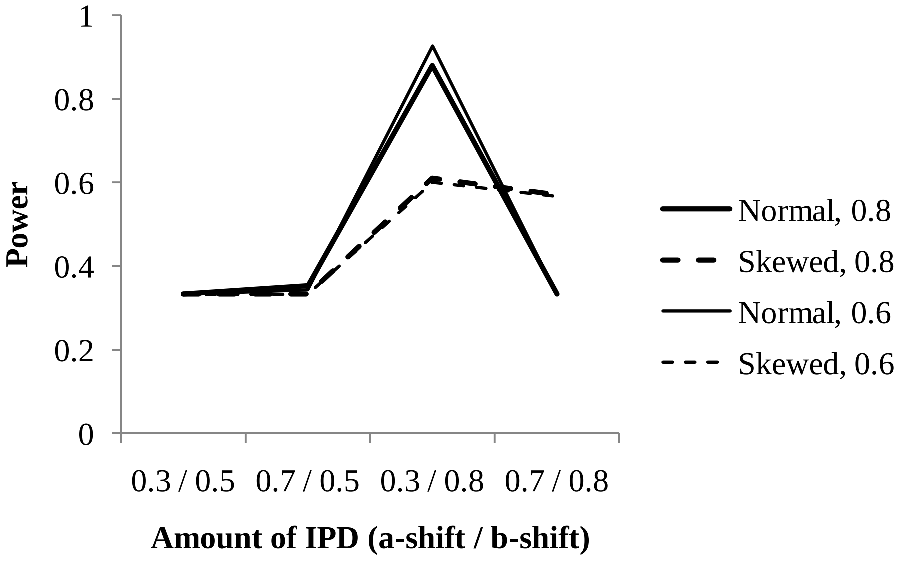


Figure 4.2. Power for Varying Levels of IPD for each Distribution Condition, 3 Spread Items, Delta Plot Method

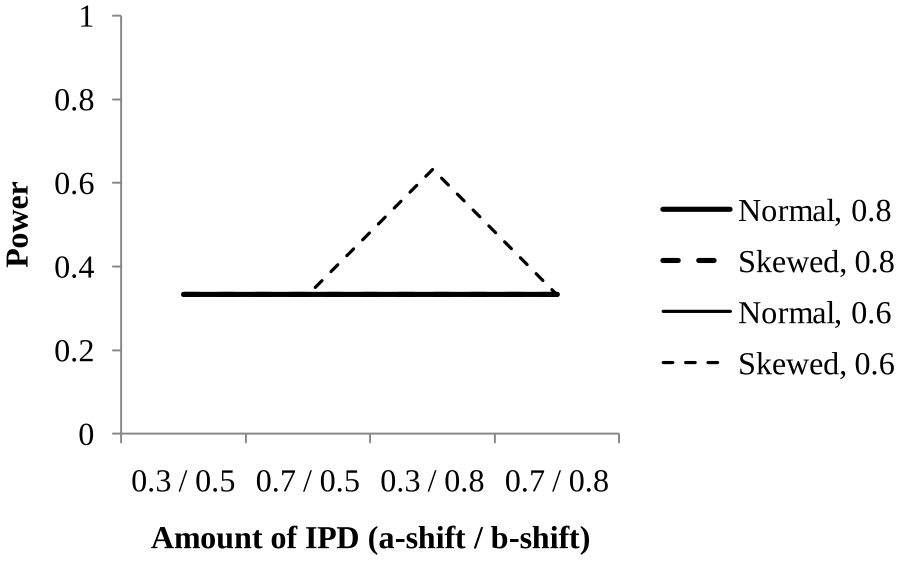


Figure 4.3. Power for Varying Levels of IPD for each Distribution Condition, 3 Moderate Items, Delta Plot Method

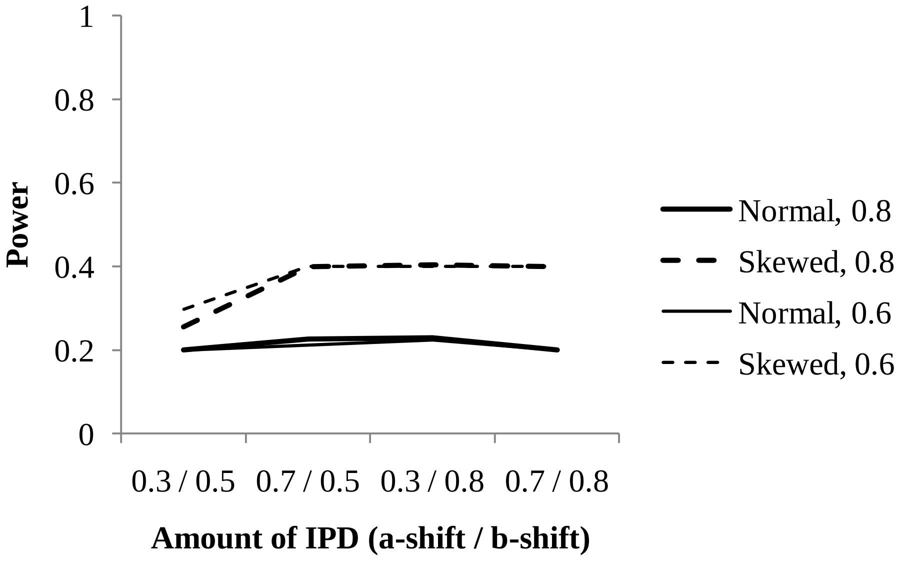


Figure 4.4. Power for Varying Levels of IPD for each Distribution Condition, 5 Spread Items, Delta Plot Method

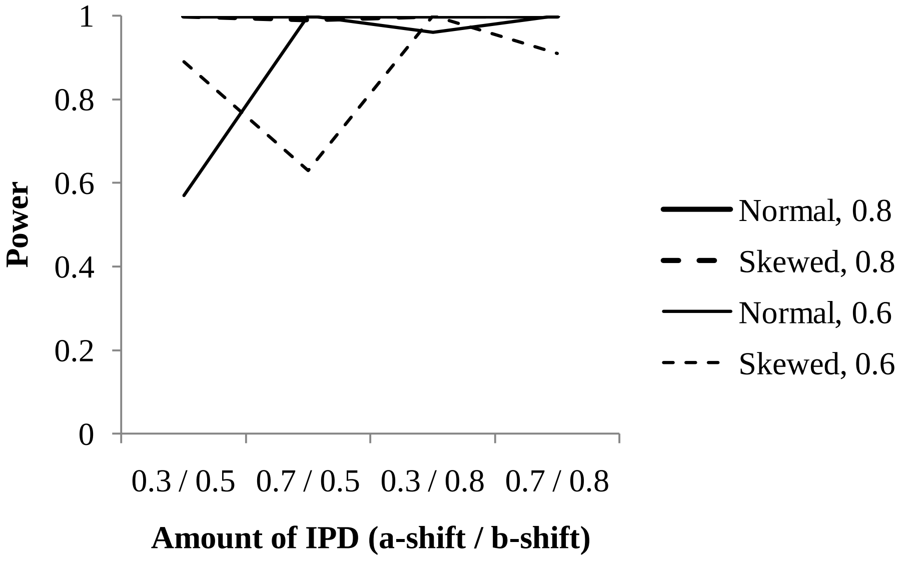


Figure 4.5. Power for Varying Levels of IPD for each Distribution Condition, 1 Hard Item, *b*-parameter plot Method

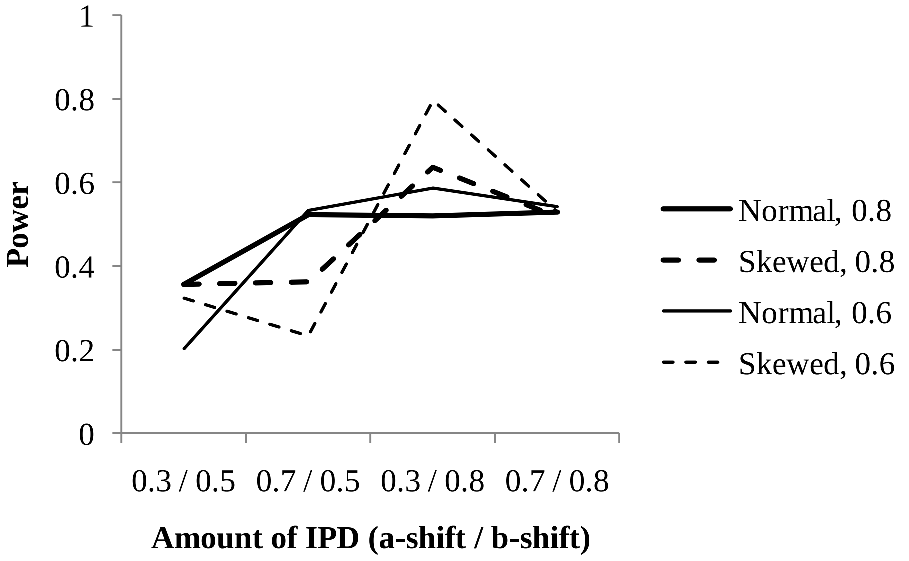


Figure 4.6. Power for Varying Levels of IPD for each Distribution Condition, 3 Spread Items, *b*-parameter plot Method

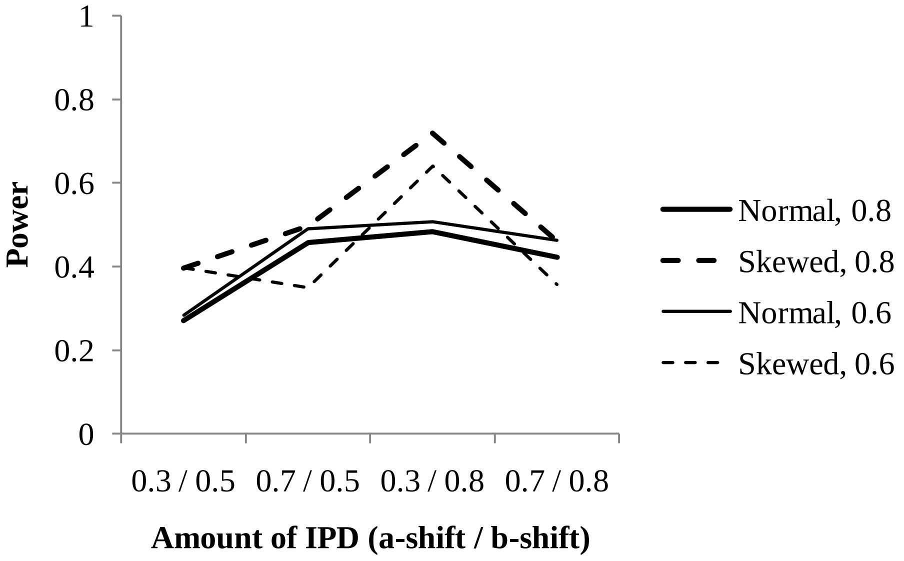


Figure 4.7. Power for Varying Levels of IPD for each Distribution Condition, 3 Moderate Items, *b*-parameter plot Method

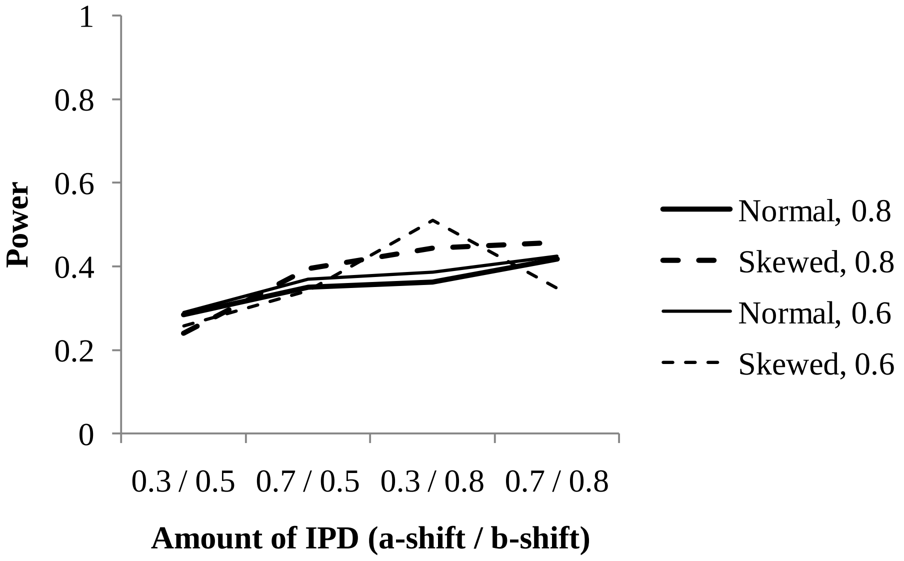


Figure 4.8. Power for Varying Levels of IPD for each Distribution Condition, 5 Spread Items, *b*-parameter plot Method

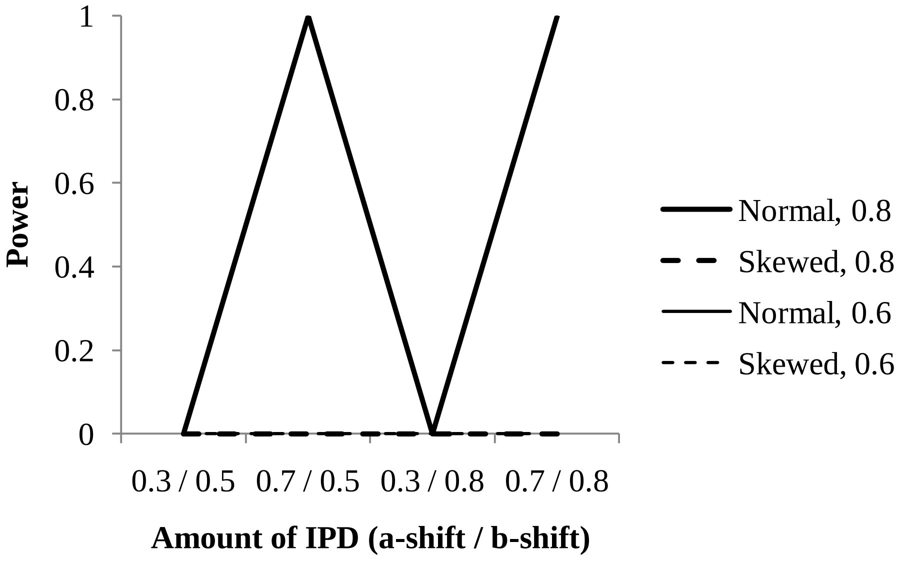


Figure 4.9. Power for Varying Levels of IPD for each Distribution Condition, 1 Hard Item, RPU Method

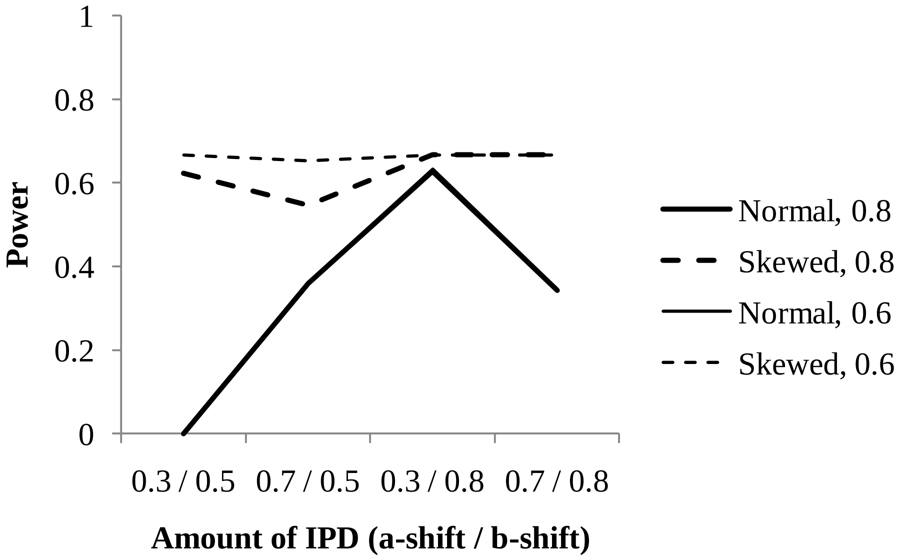


Figure 4.10. Power for Varying Levels of IPD for each Distribution Condition, 3 Spread Items, RPU Method

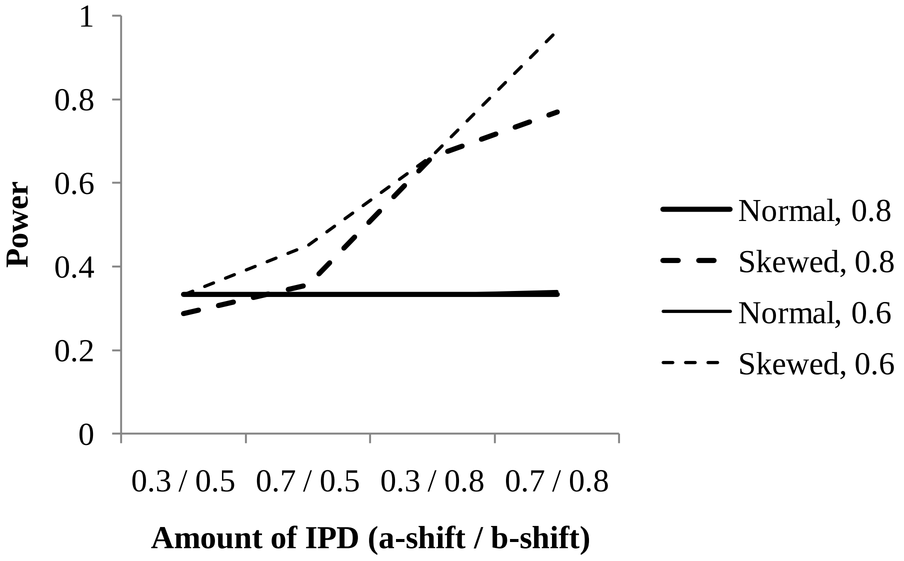


Figure 4.11. Power for Varying Levels of IPD for each Distribution Condition, 3 Moderate Items, RPU Method

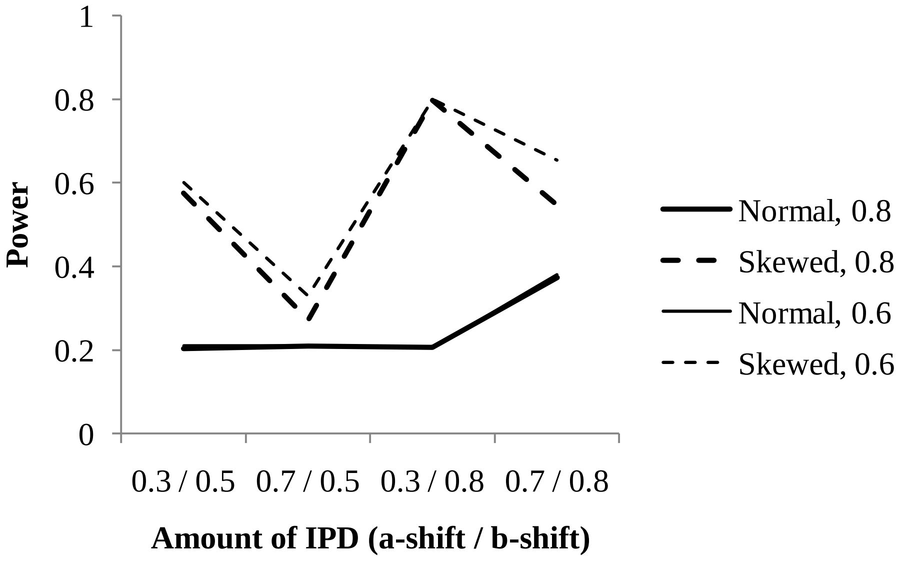


Figure 4.12. Power for Varying Levels of IPD for each Distribution Condition, 5 Spread Items, RPU Method

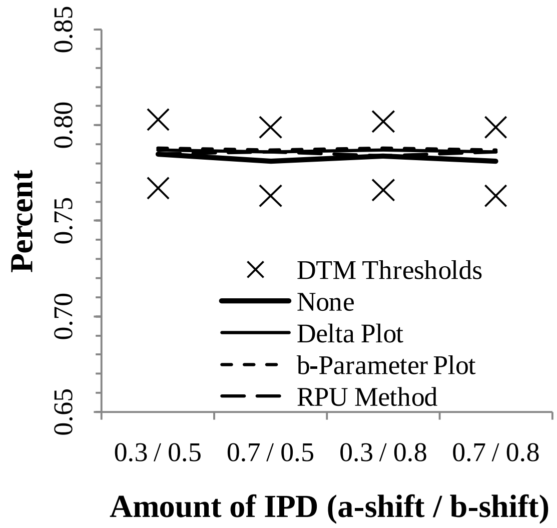


Figure 4.13. Percentage of Examinees Accurately Classified by Detection Method Employed, 1 Hard Item, No Distributional Shift, 0.8 Correlated Thetas

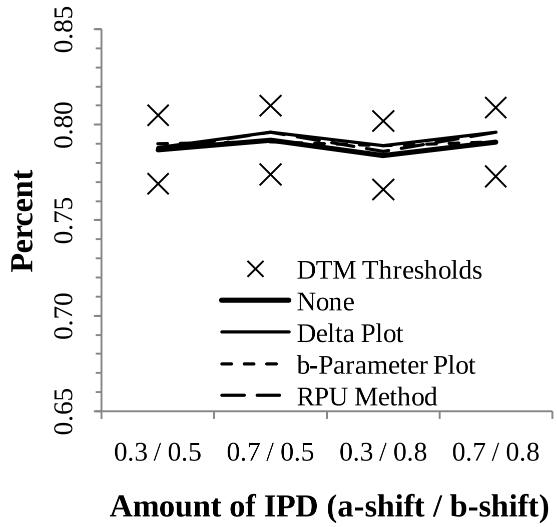


Figure 4.14. Percentage of Examinees Accurately Classified by Detection Method Employed, 3 Spread Items, No Distributional Shift, 0.8 Correlated Thetas

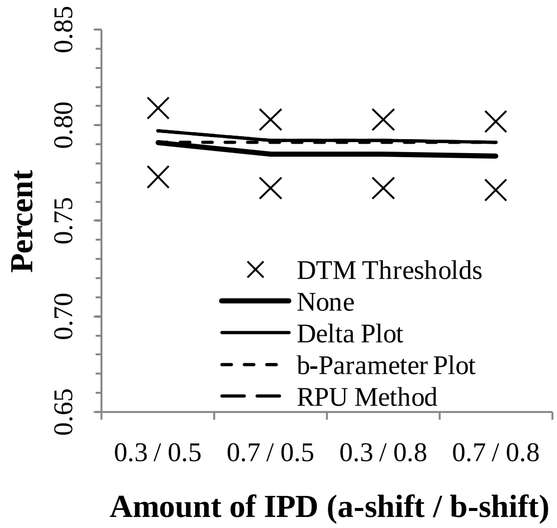


Figure 4.15. Percentage of Examinees Accurately Classified by Detection Method Employed, 3 Moderate Items, No Distributional Shift, 0.8 Correlated Thetas

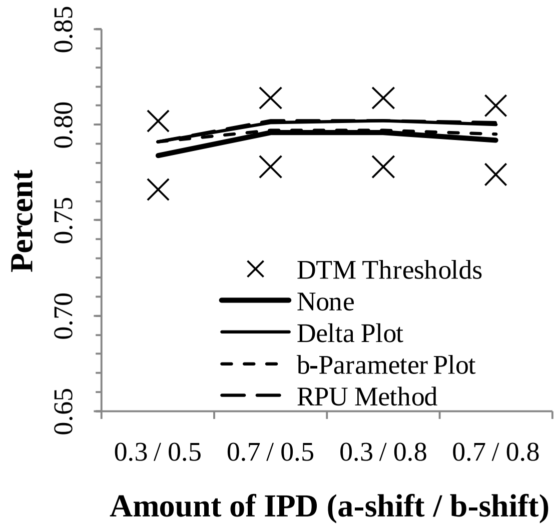


Figure 4.16. Percentage of Examinees Accurately Classified by Detection Method Employed, 5 Spread Items, No Distributional Shift, 0.8 Correlated Thetas

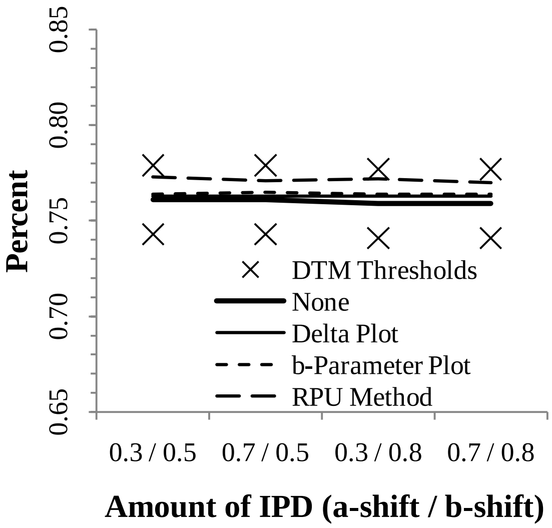


Figure 4.17. Percentage of Examinees Accurately Classified by Detection Method Employed, 1 Hard Item, Distributional Shift, 0.8 Correlated Thetas

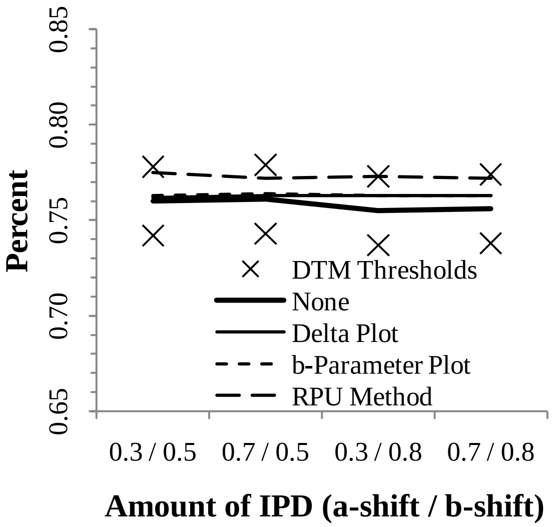


Figure 4.18. Percentage of Examinees Accurately Classified by Detection Method Employed, 3 Spread Items, Distributional Shift, 0.8 Correlated Thetas

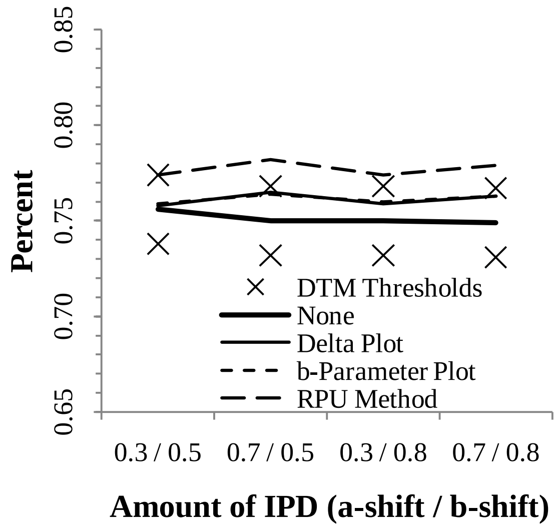


Figure 4.19. Percentage of Examinees Accurately Classified by Detection Method Employed, 3 Moderate Items, Distributional Shift, 0.8 Correlated Thetas

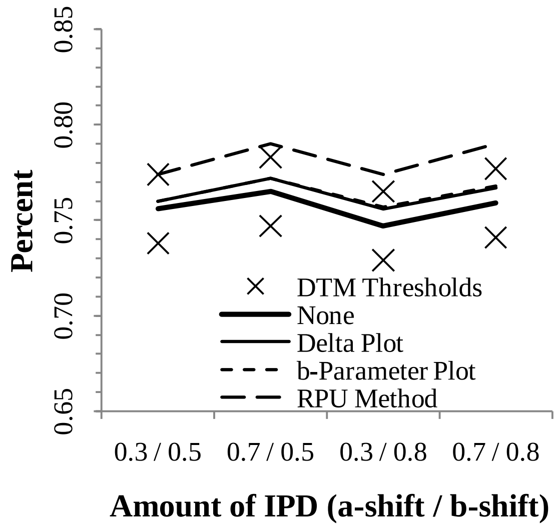


Figure 4.20. Percentage of Examinees Accurately Classified by Detection Method Employed, 5 Spread Items, Distributional Shift, 0.8 Correlated Thetas

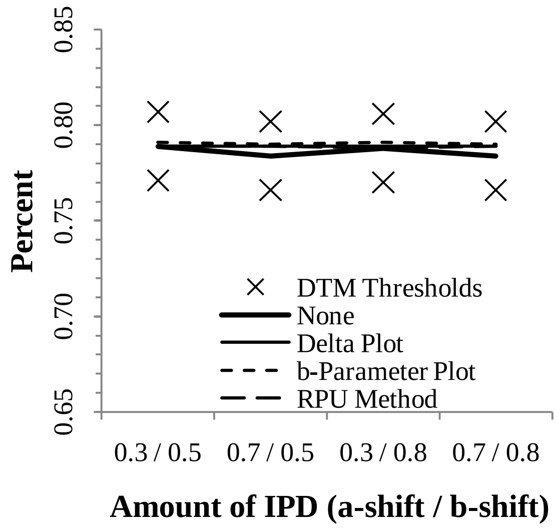


Figure 4.21. Percentage of Examinees Accurately Classified by Detection Method Employed, 1 Hard Item, No Distributional Shift, 0.6 Correlated Thetas

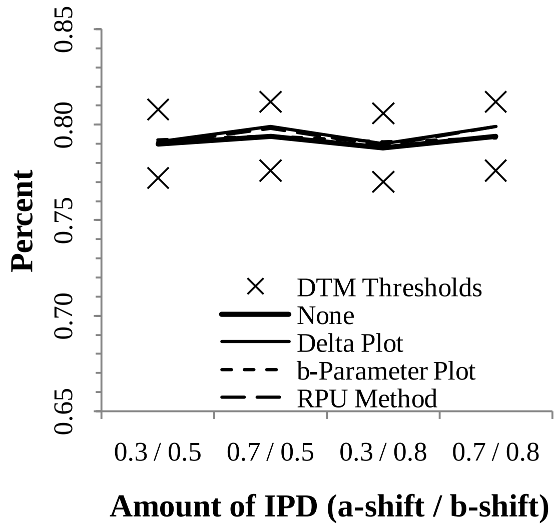


Figure 4.22. Percentage of Examinees Accurately Classified by Detection Method Employed, 3 Spread Items, No Distributional Shift, 0.6 Correlated Thetas

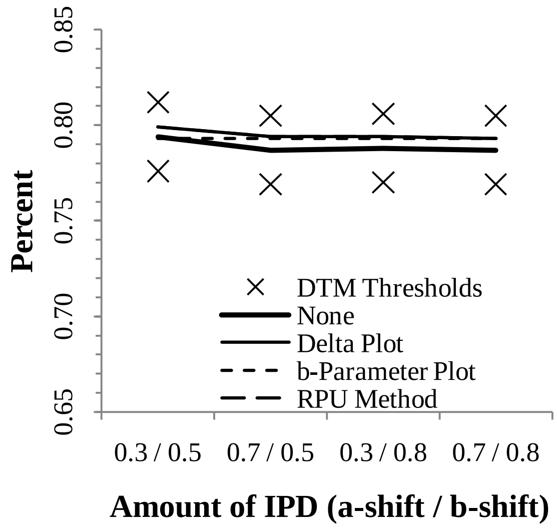


Figure 4.23. Percentage of Examinees Accurately Classified by Detection Method Employed, 3 Moderate Items, No Distributional Shift, 0.6 Correlated Thetas

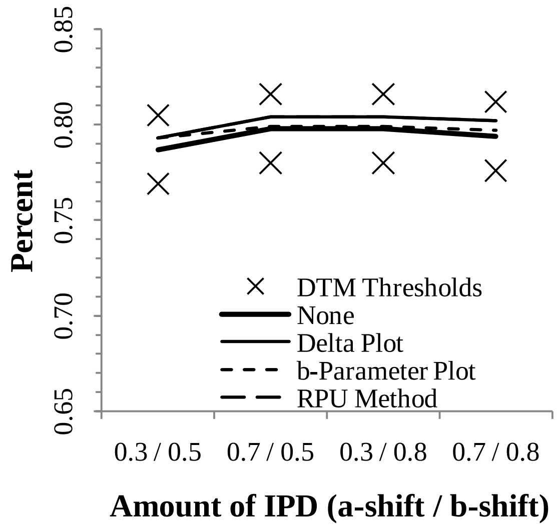


Figure 4.24. Percentage of Examinees Accurately Classified by Detection Method Employed, 5 Spread Items, No Distributional Shift, 0.6 Correlated Thetas

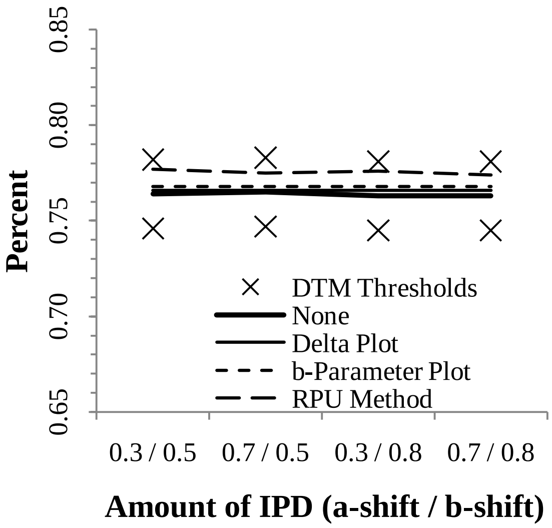


Figure 4.25. Percentage of Examinees Accurately Classified by Detection Method Employed, 1 Hard Item, Distributional Shift, 0.6 Correlated Thetas

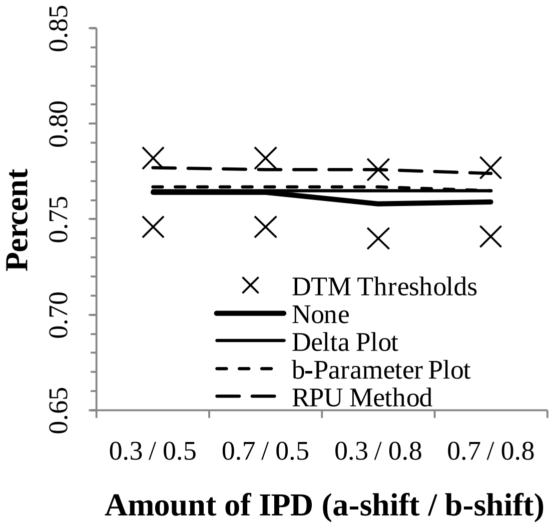


Figure 4.26. Percentage of Examinees Accurately Classified by Detection Method Employed, 3 Spread Items, Distributional Shift, 0.6 Correlated Thetas

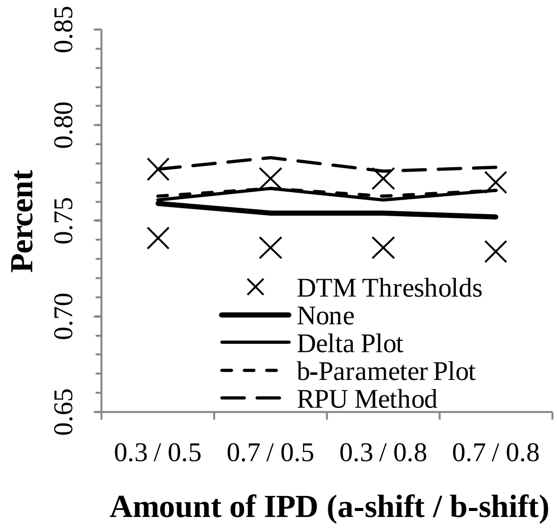


Figure 4.27. Percentage of Examinees Accurately Classified by Detection Method Employed, 3 Moderate Items, Distributional Shift, 0.6 Correlated Thetas

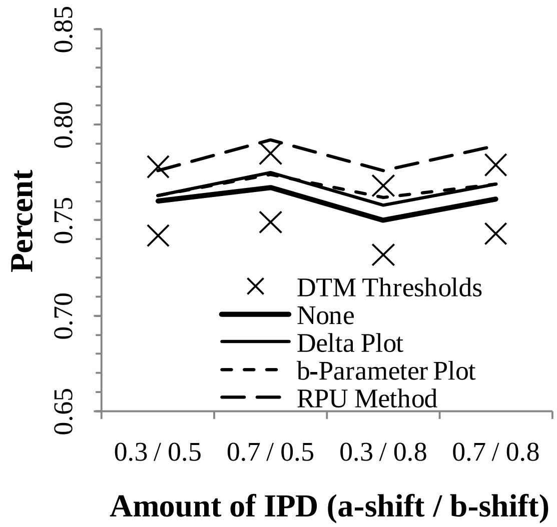


Figure 4.28. Percentage of Examinees Accurately Classified by Detection Method Employed, 5 Spread Items, Distributional Shift, 0.6 Correlated Thetas

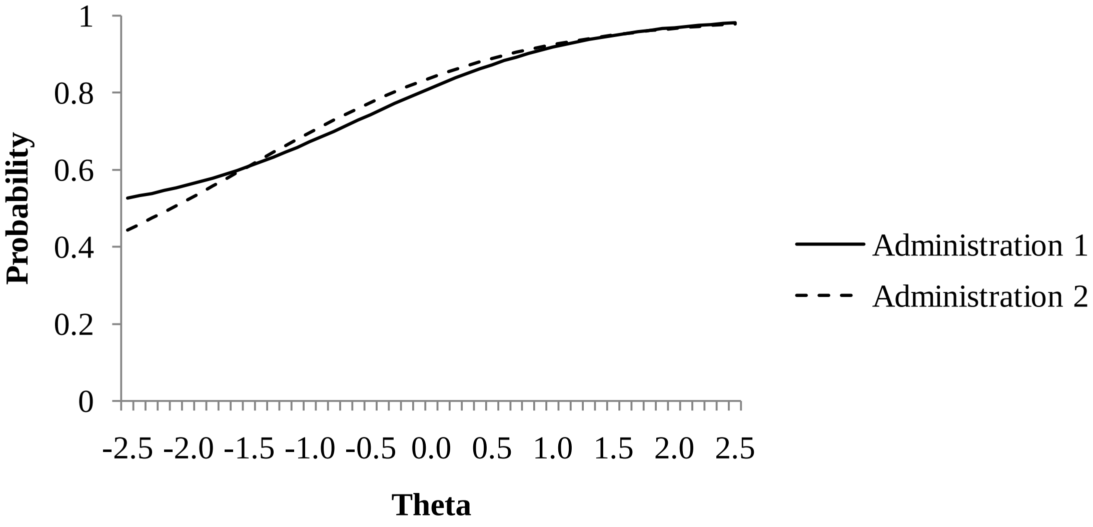


Figure 4.29. Item Characteristic Curves for Administrations 1 & 2, Grade 4, Item 12, Flagged Using *b*-Parameter Plot Method

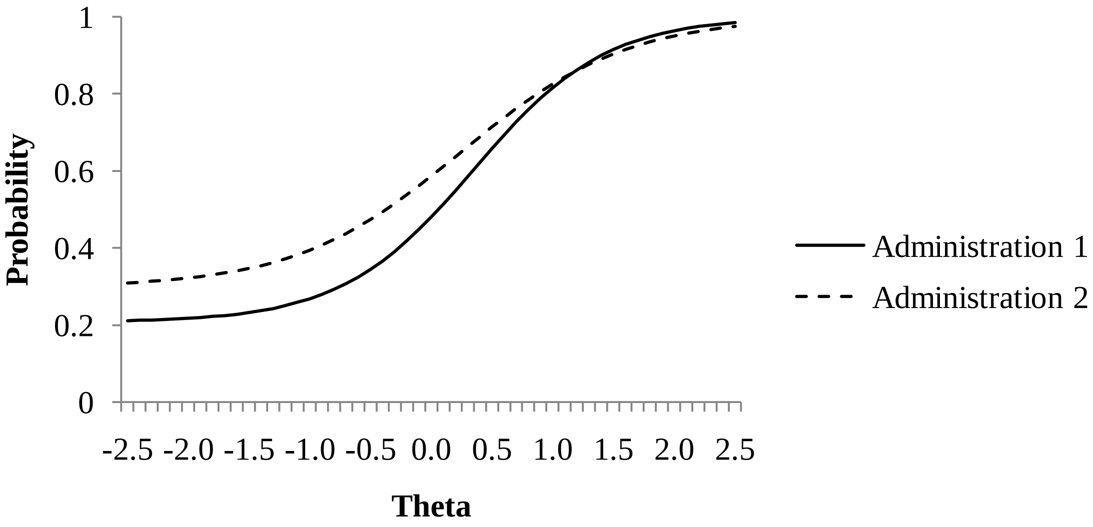


Figure 4.30. Item Characteristic Curves for Administrations 1 & 2, Grade 4, Item 22, Flagged Using Delta Plot Method

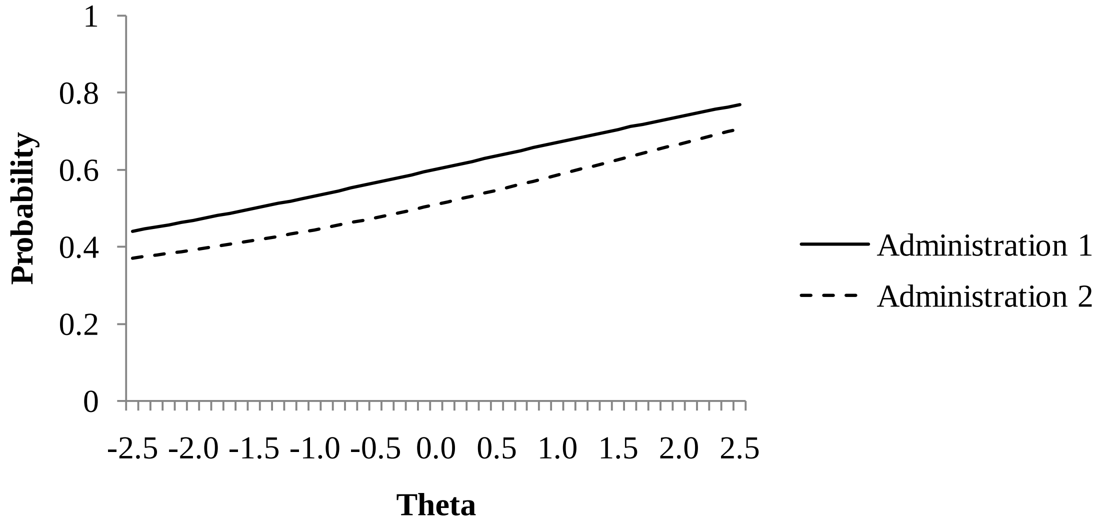


Figure 4.31. Item Characteristic Curves for Administrations 1 & 2, Grade 4, Item 26, Flagged Using Delta Plot & *b*-Parameter Plot Methods

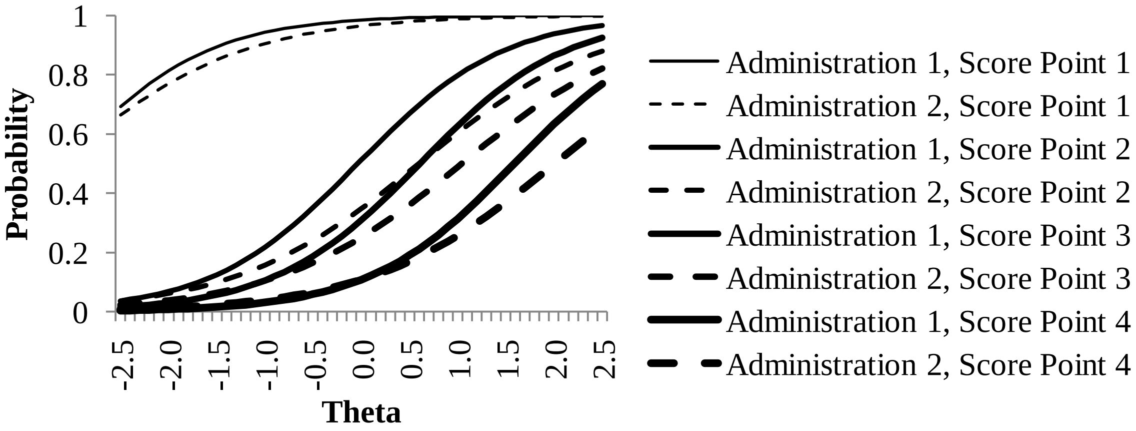


Figure 4.32. Item Category Characteristic Curves for Administrations 1 & 2, Grade 4, Item 39, Flagged Using RPU Method

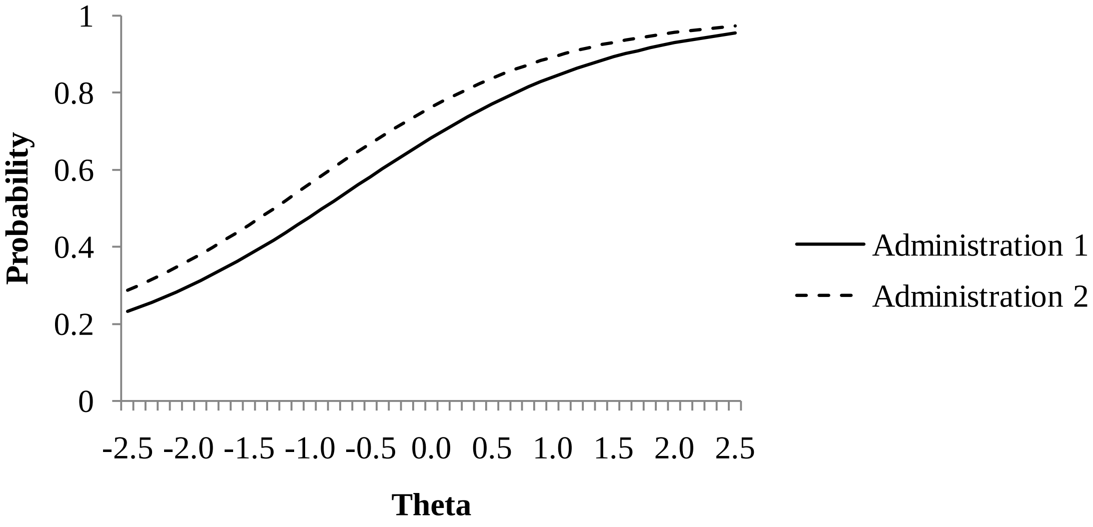


Figure 4.33. Item Characteristic Curves for Administrations 1 & 2, Grade 8, Item 6, Flagged Using Delta Plot Method

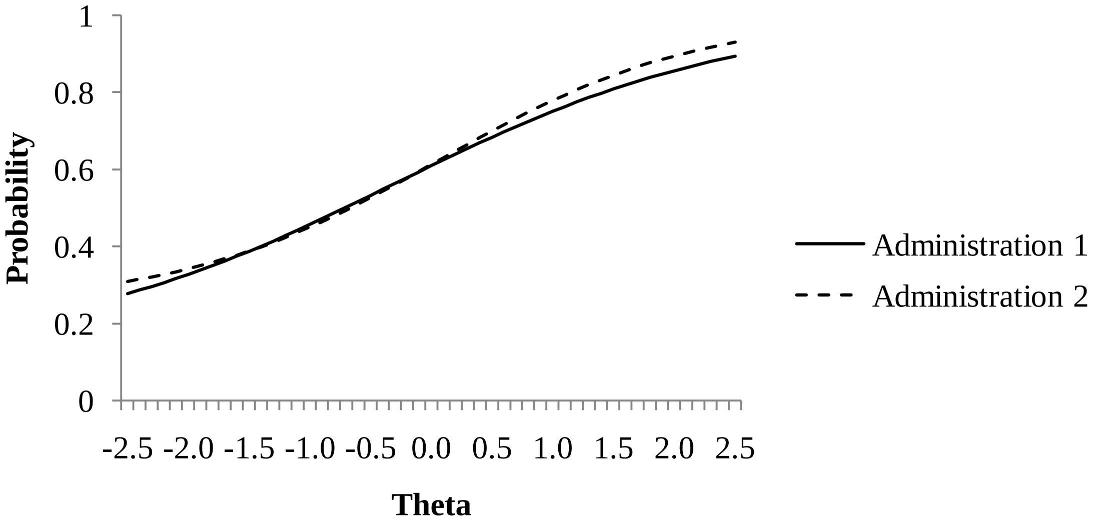


Figure 4.34. Item Characteristic Curves for Administrations 1 & 2, Grade 8, Item 17, Flagged Using *b*-Parameter Plot Method

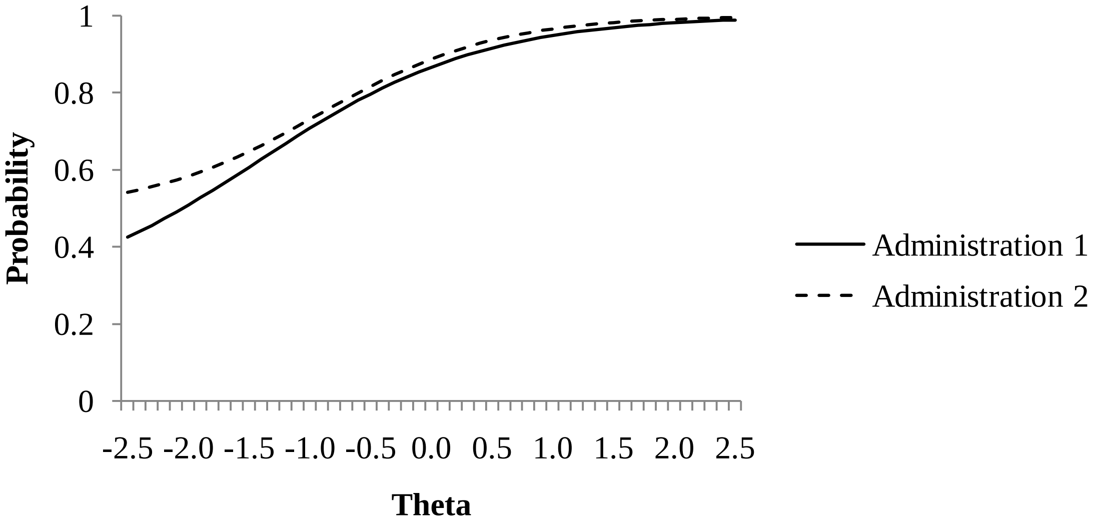


Figure 4.35. Item Characteristic Curves for Administrations 1 & 2, Grade 8, Item 22, Flagged Using *b*-Parameter Plot Method

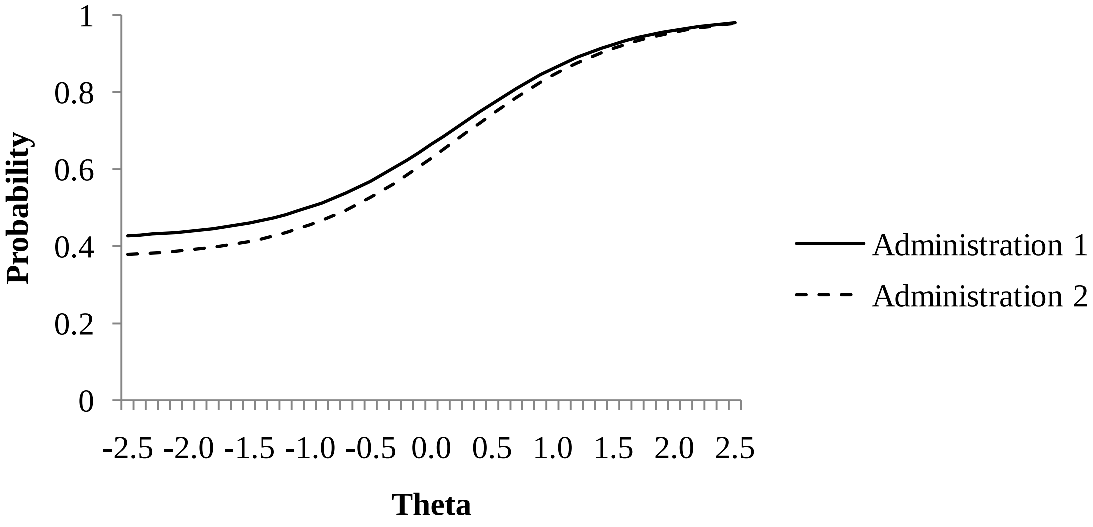


Figure 4.36. Item Characteristic Curves for Administrations 1 & 2, Grade 8, Item 28, Flagged Using Delta Plot Method

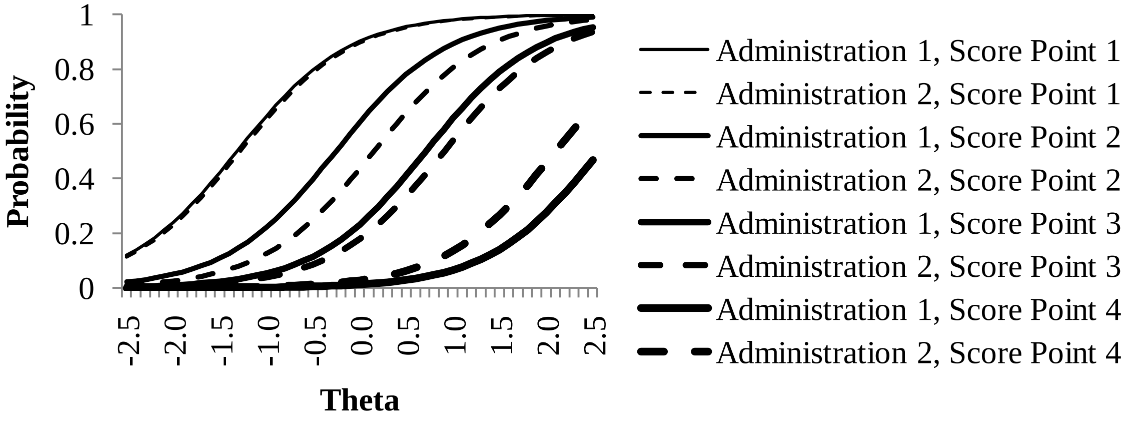


Figure 4.37. Item Category Characteristic Curves for Administrations 1 & 2, Grade 8, Item 38, Flagged Using Delta Plot Method & RPU Method

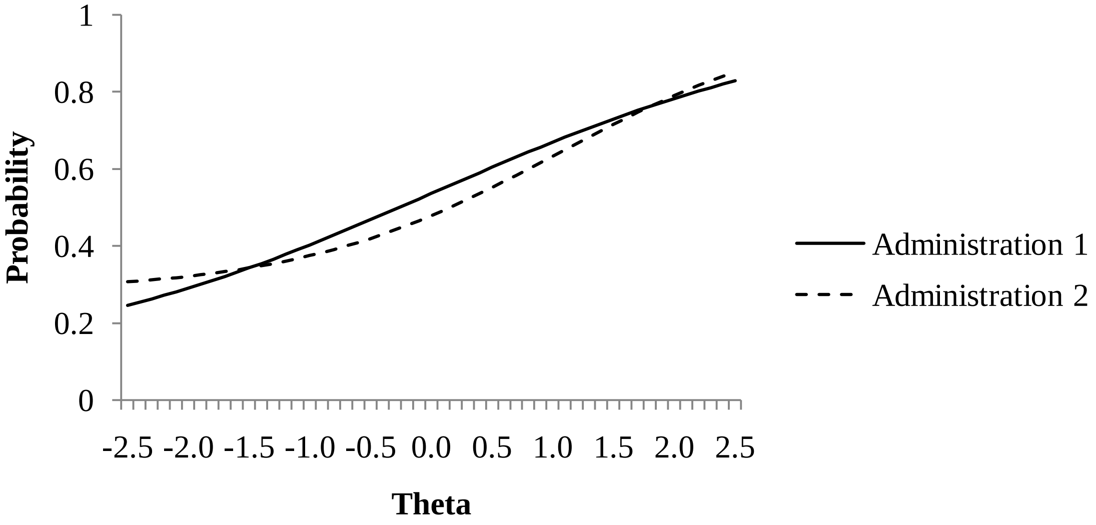


Figure 4.38. Item Characteristic Curves for Administrations 1 & 2, Grade 11, Item 3, Flagged Using *b*-Parameter Plot Method

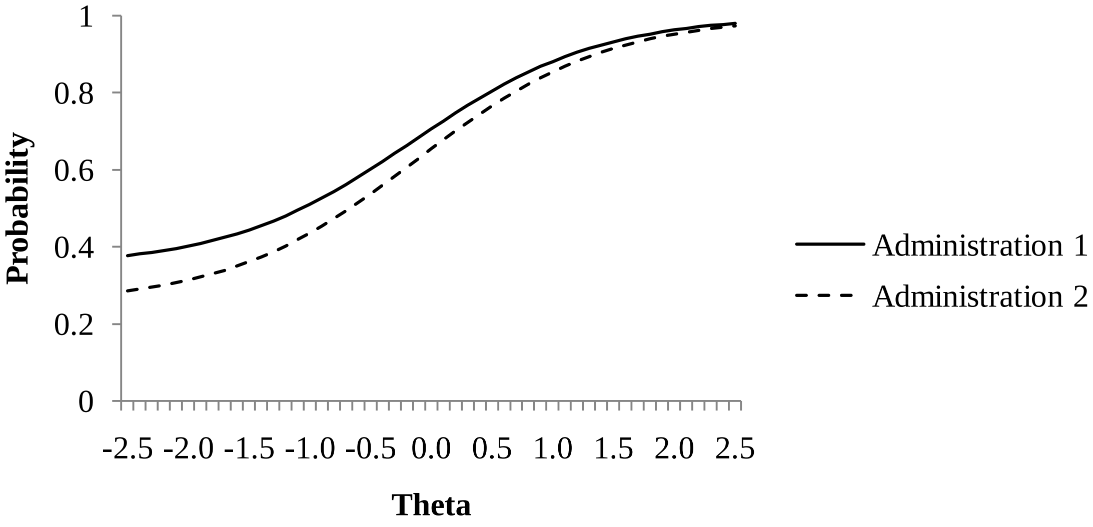


Figure 4.39. Item Characteristic Curves for Administrations 1 & 2, Grade 11, Item 12, Flagged Using Delta Plot Method

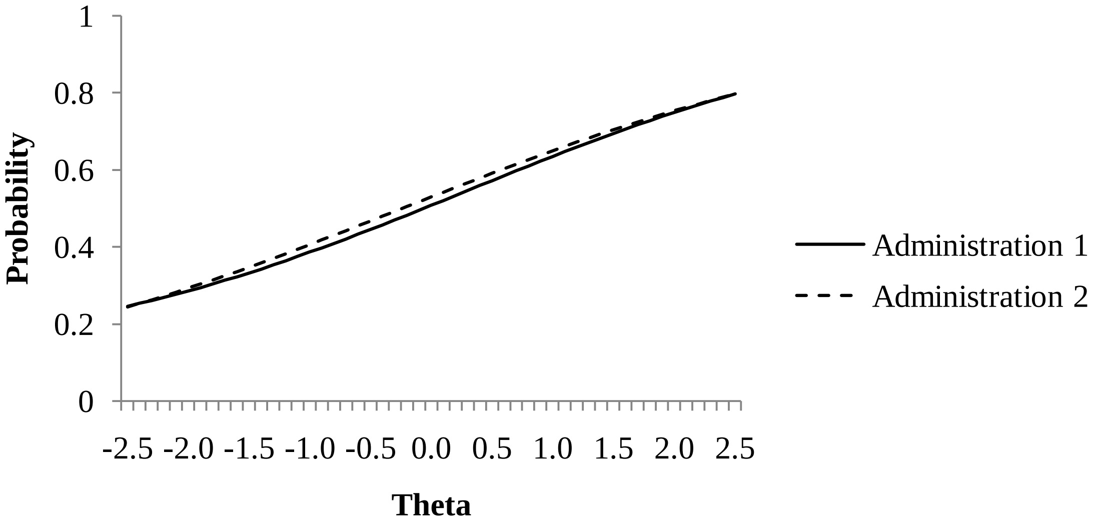


Figure 4.40. Item Characteristic Curves for Administrations 1 & 2, Grade 11, Item 14, Flagged Using *b*-Parameter Plot Method

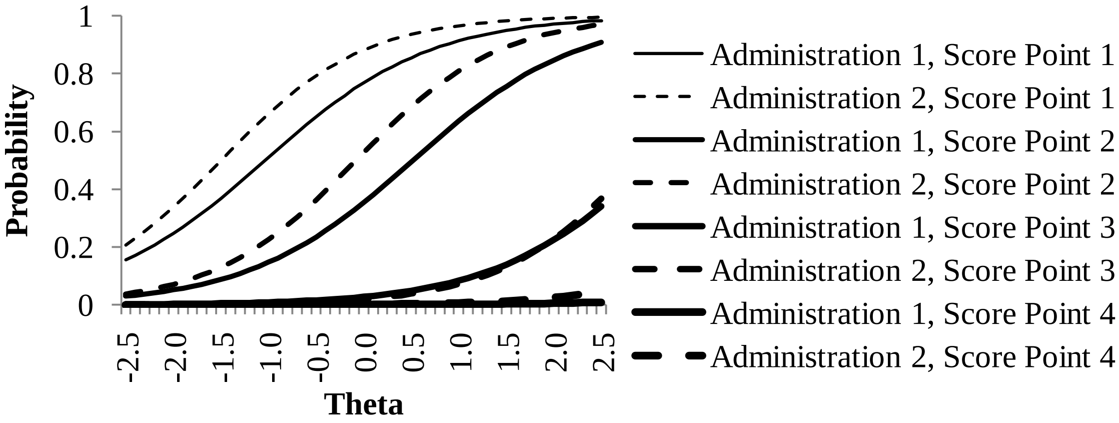


Figure 4.41. Item Category Characteristic Curves for Administrations 1 & 2, Grade 11, Item 37, Flagged Using *b*-Parameter Plot & RPU Methods

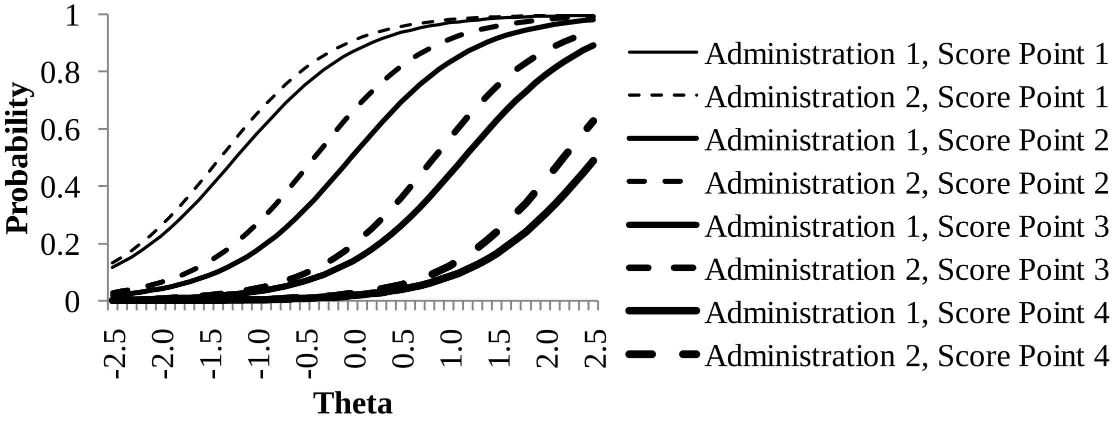


Figure 4.42. Item Category Characteristic Curves for Administrations 1 & 2, Grade 11, Item 39, Flagged Using RPU Method

CHAPTER 5

DISCUSSION

The previous chapter detailed the results of this study. This chapter serves to discuss prominent results, acknowledge limitations of the study, suggest future research directions, and present concluding remarks and recommendations for operational assessment practices.

5.1 Discussion of Findings

The primary goals for this study were to explore how potential systematic differences in opportunity to learn between assessment administrations impact the classification of examinees into proficiency categories and the ultimate representation of observed growth or lack of growth when various methods were employed for detecting item parameter drift and making anchor item removal and retainment decisions. Both simulation and empirical data analyses were conducted toward this end and a discussion of the major findings are provided in turn.

5.1.1 Simulation Study

The simulation study included five variables (i.e., administration 1 theta correlation condition, aberrant item scheme, degree of aberrancy combination, ability distribution shift condition, and aberrant item detection method) resulting in a total of 260 study conditions. Overall, too many interactions among conditions studied prevent clear interpretation of the results. Interpretations that can be summarized include the following;

(1) the RPU method was sensitive to distributional shifts between administrations resulting in inflated Type I error rates,

(2) the simulation of no drift resulted in inflated Type I error rates,

(3) average IPD detection power was poor for most cases for which more than one item was simulated with IPD,

(4) no differences in the average root mean squared error for ability parameter estimates were observed,

(5) a shift in the ability distribution between administrations resulted in lower classification accuracy,

(6) an increase in the number of aberrant items resulted in a decrease in the number of examinees under-classified,

(7) when a growth condition was simulated, the RPU method resulted both in more accurate and more under-classifications than the delta and *b*-parameter plotting methods,

(8) classification assignments were never off by more than one proficiency level,

(9) the removal of aberrant items, regardless of whether it was correctly or incorrectly flagged, and regardless of the reason for the aberrancy, resulted in more accurate classifications and fewer over-classifications of examinees,

(10) use of no detection method resulted in over estimates of growth when easier and less discriminating items appeared in the second administration, and

(11) growth rates either met or exceeded the expected values for the delta and *b*-parameter plotting methods, while the RPU method under represented the expected growth.

Further possible explanations and discussion of these findings follow.

Since multiple interactions among conditions resulted, no single IPD detection method can be recommended for practice. While all methods often struggled with power, the RPU method also struggled with a high Type I error rate when a growth condition was simulated. Thus, items were incorrectly flagged while other items were incorrectly left unflagged. Such compounding of errors resulted in practically significant differences in comparison to other methods in terms of the under- and over-classification rates of examinees. For the study conditions employed, this ultimately resulted in systematic under representations of growth.

Further, in simulating a growth condition, it was observed that Type I error rates were inflated for all detection methods when no drift was present. For both the delta and IRT *b*-parameter plot methods, this is reasonable since it is standard deviations for difference scores that are used to assess items as outliers. Thus, when no IPD exists (i.e., as in this case described), the calculated standard deviation of difference is small; thereby making it more likely for items to be flagged by chance alone. For example, in the no drift conditions, items are simulated to have identical parameters. When parameter estimates are obtained, the variation during calibration is what determines how divergent the parameters from each of the administrations for a specific item are. In this case, they should not diverge much at all. Thus, the standard deviation that results for the difference that points lie from a line of best fit is going to be quite small. Items for which it is difficult to calibrate parameters (e.g., very easy items) will be more variable and hence more likely to be flagged due to calibration errors alone and not necessarily due to IPD. This same rationale can be used to help understand the results obtained in the power analyses.

An explanation for the low power among methods, except in the case when one item was simulated as aberrant, may be that since within a condition, the degree of aberrancy was simulated uniformly (i.e., all items simulated as aberrant received the same *a*- and *b*-parameter adjustments) and it is the standard deviation of difference from a line of best fit that is used to set the flagging criteria, the accumulation of similar differences among items along with the standard error of measurement accrued during calibration of item parameters may have served to 'wash-out' the significance of the difference observed for any single item for both the delta and IRT *b*-parameter plot methods. This justification does not hold for the RPU method since the criterion used for the statistics calculated for examining IPD are not based upon the calculation of standard deviations. Rather, further explanation lies in the excessive standard error of measurement observed for some items during calibrations. Some standard errors were as high as 2.5 for the *b*-parameter estimate. This may result for items that are too easy or difficult for the examinee population and thus could certainly result in flagging errors since variations as high as these would be found to be aberrant items. This occurred for at least one item for all conditions for which the hardest item was always detected by the RPU method. Otherwise, this item was never flagged.

Despite variations in which items were flagged, no practical differences in the degree of ability parameter recovery existed between study conditions. However, the overall average difference could be consequential in cases where ability estimates border a cut score. For this study, the smallest difference (i.e., 0.75 theta points) between cut scores occurs between the second and third cuts (i.e., the most consequential as well in terms of whether an examinee contributes to meeting the threshold for making AYP). With the average RMSE being between 0.3 and 0.4 theta points, this could have significant consequences for the accurate classification of examinees into proficiency categories especially at this point. It is certainly recognized that it depends on whether examinees are located close to cut scores or not.

Likewise, very few practical differences existed in terms of the accurate classification of examinees in comparison to the case where no detection method is used for placing examinees into proficiency categories except for when a negatively skewed distribution was observed in the second administration (i.e., indicating the growth or distributional shift condition) when using the RPU method for item detection for the instances where three moderate to hard items and five spread items were simulated as aberrant. For such cases, an increase in the percent of examinees classified accurately is observed. This finding is consistent with the findings of Keller, Egan, and Schneider (2010). However, this result is a bit misleading when also considering the impact on over- and under- classification of examinees along with the reporting of growth toward a goal such as AYP.

While no practical differences were observed between using the delta plot and IRT *b*-parameter plot methods, practical differences in classifications between these methods and the RPU method did emerge for several instances where growth is simulated and ranged from 1.8 to 3.4 percent classification differences. The RPU method resulted in more under-classifications than any other method of detection employed. Likewise, fewer over-classifications were observed using the RPU method as observed also in the Keller, Egan, and Schneider (2010) study. Ultimately, the removal of an aberrant item, regardless of whether it was correctly or incorrectly flagged, and regardless of the reason for the aberrancy, resulted in more accurate classifications and fewer over-classifications of examinees suggesting that purification of the anchor results in more accurate depictions of growth.

Lastly, for all conditions, the RPU method misrepresented the amount of growth by indicating that examinee performance had not improved as much as truth would indicate. No other method under represented growth. Further, the use of no detection method was more likely to result in an over estimate of growth when items exhibiting IPD were uniformly easier and less discriminating in the second administration. The reverse is likely to be true when items become harder and more discriminating.

5.1.2 Empirical Data Analysis

The empirical data analysis consisted of re-analyzing archived statewide science achievement assessment data by applying the multiple IPD detection methods studied in the simulation. Doing so resulted in all methods flagging different items. The delta method for which the criterion of three standard deviations was used resulted in flagging no items as aberrant; and thus, growth was observed in grades 4 and 11 such that 2.9% more and 3.1% more examinees were contributing to meeting AYP, respectively. No growth was observed for grade 8 examinees.

Of all items flagged by any given alternative method and for each of the three assessments, only one item was flagged by more than one detection method. Otherwise, no more than three items were flagged for any given method and across all methods a total of four or five items were flagged. Especially worthy of note, only polytomous items were flagged when using the RPU procedures for detecting aberrant items. It was also observed that when using any other method for item detection, the percent change of examinees contributing to meeting AYP between administrations would not significantly change for grade 4. For grade 8, all methods except for the RPU method resulted in the observation of no growth between administrations. However, when using the RPU method for detecting aberrant items, a growth of 2.9% would have been observed. For grade 11, there was a practical difference in the change when the RPU method was employed such that 2.1% fewer examinees would be considered as contributing toward meeting AYP.

Inconsistencies among the detection methods suggest item removal decisions must be made with caution as it appears that methods are sensitive to different aspects that may produce IPD. For example, while the *b*-parameter plot method seemingly accounts for variations among the *b*-parameter estimates between administrations, it does not also account for potential differences in *a*- and *c*- parameter estimates that serve to produce almost identical item characteristic curves. Thus, items may be incorrectly flagged and is discussed further in the next section.

5.1.3 Linking Simulation & Empirical Study Findings

The consistency to which IPD detection methods flagged similar items within the empirical data analysis was poor, with only one item within each of the three grade levels consistently being flagged. Within the simulation study, when items were flagged the items were typically the same across IPD detection methods employed, except in the case of the RPU method where Type I errors were more prominent for the condition where growth was simulated. However, this is expected since drift was intentionally simulated for most items flagged. Additionally, for items for which no drift was simulated, the generating item parameters were identical between administrations which would rarely be the case for empirical data. Thus, they likelihood for Type I errors within simulated data conditions is surely to be smaller than for that within empirical data due to the ideal structure of the simulation design.

A more interesting finding within the empirical data analysis was that only polytomous items were flagged when using the RPU method for IPD detection. It certainly is not surprising that polytomous items were not flagged in the simulation study by any of the methods employed since drift was not simulated for these items. However, the issues that arise with the flagging of polytomous items in the empirical data analysis are multifaceted. First, what criteria should be used for flagging an item? In this study, each of the difficulty thresholds were evaluated for all IPD detection methods and an item was flagged if at least one threshold met the flagging criteria. Alternatives are available and for example, include evaluating the global difficulty parameter instead of individual thresholds. However, the greater importance here is that in the empirical data analysis, polytomous items were flagged using the criteria implemented and for some of these items, more than one IPD method flagged these polytomous items indicating that OTL may be a bigger issue within polytomous items than within multiple choice items. Further, it raises the issue of whether polytomous items should even be used as equating items since they are often even confounded by differences in rater biases. Additionally, these items are represented more heavily than multiple choice items in an assessment in terms of the number of points they contribute to scores and to the anchor set such that they have more impact when retained or removed. Thus, a decision to remove a multiple choice item from equating has a smaller impact overall than that of removing a polytomous item.

5.1.4 Relating Study Findings to Past Research

Recent past research (Karkee & Choi, 2005; Keler, Egan, & Schneider, 2010; Michaelides, 2006; Sukin & Keller, 2009b, 2010) has illuminated deficiencies associated with the delta plot method as an IPD detection method. However, the research findings presented in this simulation study do not support the use of either the *b*-parameter plot method or the RPU method over the delta plot method. However, retaining items for equating that are decidedly aberrant also is not supported by this research, regardless of the cause of the aberrancy. Even while these methods did not flag all items that perform aberrantly, the removal of those that are flagged proved to produce more accurate classifications of examinees and also proved to provide closer depictions of the amount of growth occurring between cohorts of examinees. The increased accuracy of classifications through the use of detection methods is supported by previous research as well (e.g., Keller & Wells, 2009).

However, while the only research on the RPU method (Keller, Egan, & Schneider, 2010) as an IPD detection method presents promising results, this research has found differing results suggesting that the RPU method may be more sensitive to changes in cohort ability distributions than first thought. However, this finding may have occurred in combination with a unique component of this research which is that the response patterns for the first administration were simulated with two degrees of multidimensionality in an attempt to highlight impacts on evaluation criteria as it is related to a lack of opportunity to learn a specific content area. Thus, the dimensionality of the assessment was manipulated between administrations in such a way to represent a theory of improved instruction based on past test results such that the lack of opportunity to learn disappears in the second administration. Thus, response patterns in the second administration were simulated as unidimensional. Therefore, the differences between the Keller, Egan, and Schneider (2010) research study and this one may be directly related to this difference in study design, suggesting that the RPU method is sensitive to dimensionality differences. Further, more liberal flagging criteria was employed in this research design which may further highlight reasons for the differences found between these studies since whenever more liberal flagging criteria is employed, the likelihood for Type I errors increase as a result.

5.1.5 Implications of Differences in Opportunity to Learn

The broader implications of the findings presented in this work relate directly to answering the question of "Do differences in opportunity to learn mask growth if aberrant items are removed and the cause of the aberrancy is directly related to instructional efforts for that particular content strand?" In responding to this question, it is helpful to illuminate once more how OTL was simulated in this study. Within the first administration of the simulation study, item response data was simulated under conditions of bi-dimensionality such that one dimension represented the construct competency of examinees assuming the specific areas (i.e., content strands) of the construct were sufficiently taught and the other dimension represented the lack of opportunity to learn other specific areas of the construct due to the negligence to teach that content strand. Correlations between the two dimensions were assumed to still likely be quite high since regardless of the specification of content strands, the construct is still considered unidimensional. Thus, even while examinees may not be instructed in a particular strand, knowledge from other areas may aid the examinee in responding correctly to items from neglected strands. The amount of correlation between these two dimensions was simulated both as 0.6 and 0.8 to determine (1) whether the magnitude of the correlation between the construct and OTL impacts the outcome variables and (2) to determine if as an assessment becomes unidimensional (i.e., through a more thorough coverage of the construct in instruction) whether true growth is masked by the removal of items that exhibit IPD from the anchor that is used to equate test forms. Results show that the differences between the magnitudes do not have a great impact on the resulting outcome variables except in a few cases as summarized in Section 4.2.5.1. Further, growth does not seem to be masked, but rather if items are removed from the anchor, under the conditions of a matrix-sampled design, both the accurate classification of examinees and the estimate for reported growth toward AYP improves. Otherwise, maintaining the offending items resulted in more classification errors and inflation within the amount of growth to be reported.

However, the choice of detection method matters in the sense that methods resulting in Type I errors are a detriment to the accurate classification of examinees and thus, accurate depictions of growth. This is exemplified using the RPU method when growth is simulated between administrations. Further, under a condition of no drift, Type I errors were also prominent indicating that differences in dimensionality between assessments may impact the proper calibrations of item parameter estimates. This results from forcing a unidimensional calibration method onto multidimensional data. Thus, the resulting calibrations extract what is common between items. What is common is intended to be the construct of interest. However, in the presence of confounding factors such as opportunity to learn, what is common between items is reduced and the resulting calibrations do not fully represent the construct in its operational entirety. It is not until differences in opportunity to learn are removed that calibrations can become more invariant across administrations. Regardless, it does seem that so long as an effort is made to detect and remove aberrant items using a method that is known to produce few Type I errors, growth between cohorts can be approximated well.

Since growth approximations between cohorts seemed to improve with the removal of correctly identified aberrant items, it seems that even in the presence of OTL differences, examinee classifications are robust. However, to the extent that detection methods are sensitive to Type I error, the resulting equated scores are more likely to be flawed. Thus, it could be argued that more conservative flagging criteria be employed since the combination of poor power and controlled Type I error rates seemed to produce growth determinations closer to expected values.

Overall, it can be said that there is an interaction between the assessment of a teachers' breadth of instruction and a students' breadth of knowledge in relation to the material tested. To the extent that this matters depends on how much teachers collectively diverge from the tested academic standards. In the study findings presented here, based upon a matrix-sampled equating design, examinee classifications were robust to differences in OTL between cohorts. This is likely due to the strength of the equating design employed. Within the matrix-sampling design, a large of number of items are used for equating test scores which increases the reliability of the adjustments. Additionally, even while items within a specific content strand may be more likely to be removed due to detected difference in OTL between administrations, examinees are still scored using items that represent these content areas and thus the content was unidimensional enough (i.e., the two ability parameters from the first administration were correlated enough) to still capture depictions of growth that approached truth. This robustness seems to dissipate as polytomous items are questioned as these items carry more weight than a single multiple choice item. To the extent that empirical data mimics the simulation design outlined, difference in OTL between cohorts does not seem to call into question decisions made about the growth between cohorts. Conversely, to the extent that the polytomous items were correctly flagged using the RPU method within the empirical data presented and those items are divergent between administrations due to differences in OTL and instructional emphasis for the content addressed, difference in OTL between cohorts seems to significantly call into question decisions made about growth between cohorts.

5.2 Limitations of the Study

A multitude of limitations exist for this study and the most concerning will be discussed within this section. First, all IPD simulated within any one condition was both of the same magnitude for all items and in the same direction. This is unlikely to occur in empirical data. It is more likely that while some items are calibrated as easier, other may too be calibrated as harder for a variety of causes (e.g., de-emphasized instruction in a particular content strand or more stringent scoring of constructed response items). Additionally, only two administrations were investigated and it is also likely that as IPD goes undetected, it becomes more pronounced over time and practical consequences for this are left unexplored in this research. Further, the comparison of ability parameter estimates to those simulated as truth may be an unrealistic and contaminated comparison. Rather, the comparison with the resulting ability parameter estimates from a condition of no simulated drift may be better to account for the inherent error associated with calibrating parameter estimates. Another element that limits the generalizability of the simulation results to empirical data is the way in which multidimensionality was simulated. Conditions of 0.8 and 0.6 correlations between ability parameters (i.e., two dimensions were simulated) were employed for the first administration and unidimensionality was maintained for the second set of ability parameters intended for the second administration. These magnitudes of correlation are not known to be common within empirical data and the cleanliness to which dimensionality was simulated is certainly not reflective of empirical data where the opportunity to learn may be more unevenly dispersed than simulated.

5.3 Directions for Further Research

Since it was common among both the delta and IRT *b*-parameter plot methods that only one item was flagged within the simulation study, resulting in poor average power when more than one item was simulated as aberrant, it may be wise to investigate both methods in an iterative manner, such that items exhibiting IPD be removed and then the standard deviation of the difference for remaining items be calculated and used as the new criterion for item removal decisions. Upon the subsequent iterations, this would serve to create a more conservative criterion and thus increase the likelihood that aberrant items would be flagged. Thus, this procedure would involve the extra step of updating the criterion upon item flagging and reassessing until no items are flagged as aberrant. However, it too is expected that more Type I errors would result. As observed for the control condition in the simulation study when no aberrancy was simulated, items were still detected as exhibiting IPD at higher rates than when items were simulated with drift due to the reduction in the standard deviation of the difference from the line of best fit and also due to the overlap of item parameter calibration error. Thus, further research is needed before implementing such iterative practices.

It may also be useful to explore various flagging criteria as the preliminary analysis for this study resulted in the decision to invoke more conservative criteria. However, this was done based on common implementation of methods and not in any systematic way. Invoking a systematic exploration of magnitudes for criteria under varying conditions may be warranted.

While the likelihood ratio test is difficult and time consuming to implement in simulation studies, overcoming these barriers would serve the field well as previous preliminary work has been promising (Sukin & Keller, 2009b, 2010) and may provide a better detection method that would not be as difficult or time consuming to implement when used in practice (i.e., a vast number of multiple replications would not be required). Finally, it was interesting that the RPU method flagged primarily constructed response items. Therefore, it may be that OTL differences are manifested within this item type and their use as equating items needs further investigation.

5.4 Conclusions & Recommendations

The importance of this study is clear in regards to the current large-scale educational assessment climate, where the accurate classification of students into performance categories is essential. Additionally, with assessments influencing the content of instruction, the likelihood of finding outlying anchor items is high and so deciding how to appropriately detect and deal with these items is of the utmost importance.

Ultimately, the accumulation of results based on both the simulation study and the empirical data analysis provide support for eliminating the use of flagged items for linking assessments when a matrix-sampling design is used and a large number of items are used within that anchor. Of course, it is always recommended that content specialists be consulted before removing items as maintaining content balance within the anchor test is important, especially when reporting subscale scores. It is further recommended that the RPU method be further researched before being recommended for use in practice due to its higher Type I error rates in the presence of a distributional shift and its practically significant classification differences exhibited both within the simulation and empirical data analyses. Further, while neither the delta nor the IRT *b*-parameter plot methods produced results that would overwhelmingly support their use, it is recommended that both methods be employed in practice until further research is conducted for alternative methods. Alternatively, use of delta plots and the RPU method supported with subjective analyses of ICCs and ICCCs for flagged items may also serve as sound anchor purification practice. However, like any new suggestions and ideas for a change in practice, more research is needed to both corroborate the findings of this work and follow-up on suggested procedures. It may be possible within operational settings to build this line of research by conducting both traditional analyses and supplementing those traditional analyses with newer procedures to explore differences in how results would be reported, much like the work presented here within the experimental data analysis.

APPENDIX A

SUPPORTING TABLES AND FIGURES

Table A.1. Alphabetical Reference Guide to Acronyms

Acronym	Description
2-PLM	Two-Parameter Logistic Model
3-PLM	Three-Parameter Logistic Model
AYP	Annual Yearly Progress
CDIF	Compensatory Differential Item Functioning
CINEG	Common Items Non-Equivalent Groups
CR	Constructed Response
CTT	Classical Test Theory
DFIT	Differential Functioning of Items and Tests
DIF	Differential Item Functioning
DTM	Difference That Matters
ESEA	Elementary and Secondary Education Act
ETS	Educational Testing Service
FCIP	Fixed Common Item Parameter
GRM	Generalized Response Model
HB	Haebara
ICC	Item Characteristic Curve
ICCC	Item Category Characteristic Curve
ICTC	Item Category Threshold Curve
IPD	Item Parameter Drift
IRT	Item Response Theory
LR	Likelihood Ratio
MC	Multiple Choice
MH	Mantel-Haenszel
MM	Mean-Mean
MS	Mean-Sigma
NAEP	National Assessment of Educational Progress
NCDIF	Non-Compensatory Differential Item Functioning
NCLB	No Child Left Behind
NEAT	Non-Equivalent groups Anchor Test
OTL	Opportunity to Learn
RMSE	Root Mean Squared Error
SA	Short Answer
SD	Standard Deviation
SL	Stocking and Lord
TCC	Test Characteristic Curve
TID	Transformed Item Difficulty

Table A.2. MIRT Generating Parameters, Administration 1

		Parameter							
	Form #	*c*	*a1*	*a2*	*γ*	*b1*	*b2*	*b3*	*b4*
Scoring Items	All	0.30	1.13	0.00	-0.16				
	All	0.26	0.44	0.00	-0.20				
	All	0.25	0.29	0.81	-0.50				
	All	0.25	1.00	0.00	0.14				
	All	0.24	0.56	0.00	-0.47				
	All	0.23	0.89	0.00	-0.12				
	All	0.22	1.11	0.00	-1.20				
	All	0.22	0.79	0.00	-0.13				
	All	0.21	0.99	0.00	-0.45				
	All	0.20	0.72	0.00	0.11				
	All	0.19	1.06	0.00	-1.00				
	All	0.18	0.00	1.07	-1.63				
	All	0.18	0.70	0.00	0.27				
	All	0.17	0.74	0.00	0.48				
	All	0.17	0.46	0.52	0.10				
	All	0.17	0.49	0.00	-0.47				
	All	0.16	0.77	0.00	0.51				
	All	0.15	0.32	0.00	0.07				
	All	0.15	0.99	0.00	-0.08				
	All	0.14	1.23	0.00	0.90				
	All	0.14	0.69	0.00	0.15				
	All	0.14	1.21	0.00	0.81				
	All	0.13	0.35	0.00	0.80				
	All	0.13	0.46	0.00	-0.88				
	All	0.13	0.43	0.41	-0.02				
	All	0.11	0.82	0.00	0.81				
	All	0.10	0.46	0.00	0.45				
	All	0.08	0.79	0.00	1.72				
	All	0.08	0.57	0.00	-0.14				
	All	0.07	0.61	0.00	0.52				
	All	0.07	0.68	0.00	-0.03				
	All	0.00	0.53	0.00	0.98				
	All	0.00	0.50	0.00	0.79				
	All	0.00	0.62	0.00	-1.12	0.85	-0.85		
	All	0.00	0.85	0.00	0.34	1.55	-1.55		
	All	0.00	0.50	0.00	0.09	1.49	-1.49		
	All	0.00	0.64	0.00	0.22	2.95	-0.19	-2.76	
	All	0.00	0.74	0.00	-1.22	2.48	-0.08	-2.40	
	All	0.00	0.78	0.00	-1.23	1.71	0.01	-1.72	
	All	0.00	0.77	0.00	0.61	1.40	0.32	-1.72	
	All	0.00	0.93	0.00	0.04	2.46	0.72	-0.79	-2.39
	All	0.00	0.67	0.00	-0.95	3.54	0.73	-1.15	-3.12
	All	0.00	0.90	0.00	-1.47	2.23	0.81	-0.72	-2.31
	1	0.13	0.51	0.00	-0.30				
	1	0.12	0.59	0.00	1.07				
	1	0.10	0.60	0.00	0.91				
	1	0.12	0.61	0.00	0.94				
	1	0.24	0.72	0.00	-0.34				
	1	0.19	0.98	0.00	-0.86				
	1	0.20	1.07	0.00	0.08				
	1	0.00	1.01	0.00	-0.52	1.02	0.33	-0.35	-1.00

Continued, next page.

Table A.2., cont'd.:

Table A.2. MIRT Generating Parameters, Administration 1

					Parameter				
	Form #	*c*	*a1*	*a2*	*γ*	*b1*	*b2*	*b3*	*b4*
Linking Items Administration 2 No Parameter Shifts	2	0.10	0.34	0.00	-0.08				
	2	0.24	0.46	0.00	-0.49				
	2	0.20	0.48	0.00	0.83				
	2	0.15	0.50	0.00	0.42				
	2	0.13	0.70	0.00	0.49				
	2	0.42	0.87	0.00	-0.18				
	2	0.27	0.87	0.00	-0.66				
	2	0.27	0.92	0.00	-0.66				
	2	0.30	1.02	0.00	-0.97				
	2	0.00	1.01	0.00	-0.28	2.25	0.72	-0.74	-2.23
	3	0.18	0.53	0.00	-0.35				
	3	0.21	0.58	0.00	-0.82				
	3	0.26	0.61	0.00	0.06				
	3	0.06	0.64	0.00	0.19				
	3	0.13	0.66	0.00	0.29				
	3	0.15	0.78	0.00	0.95				
	3	0.22	0.79	0.00	0.40				
	3	0.19	0.82	0.00	0.33				
	3	0.00	0.99	0.00	-0.43	1.74	0.68	-0.28	-2.15
	4	0.12	0.42	0.00	0.13				
	4	0.39	0.50	0.00	-0.11				
	4	0.19	0.61	0.00	-0.38				
	4	0.24	0.63	0.00	0.91				
	4	0.14	0.65	0.00	0.51				
	4	0.06	0.65	0.00	0.49				
	4	0.11	0.79	0.00	0.69				
	4	0.00	0.83	0.00	-0.66	2.15	0.66	-0.81	-2.00
Linking Items Administration 2 Parameter Shifts	1	0.34	0.28	0.81	-0.45				
	1	0.14	0.74	0.65	0.55				
	3	0.27	0.41	0.85	-0.23				
	4	0.14	0.06	0.75	-1.43				
	4	0.09	0.72	0.25	0.51				

Table A.3. IRT Generating Parameters, Administration 2, Conditions 0-16

						Parameter			
	Condition Description (*a-shift, b-shift*)	Form #	*c*	*a*	*b*	*b1*	*b2*	*b3*	*b4*
Scoring Items	▨	All	0.08	0.79	-2.17				
		All	0.13	0.35	-2.30				
		All	0.21	0.99	0.45				
		All	0.16	0.77	-0.67				
		All	0.26	0.44	0.46				
		All	0.00	0.53	-1.86				
		All	0.15	0.32	-0.21				
		All	0.30	1.13	0.14				
		All	0.17	0.74	-0.64				
		All	0.10	0.46	-0.99				
		All	0.22	1.11	1.08				
		All	0.13	0.46	1.89				
		All	0.17	0.70	-0.14				
		All	0.24	0.56	0.83				
		All	0.14	1.23	-0.73				
		All	0.18	1.07	1.52				
		All	0.00	0.50	-1.56				
		All	0.23	0.89	0.14				
		All	0.18	0.70	-0.39				
		All	0.13	0.60	0.04				
		All	0.25	0.86	0.59				
		All	0.19	1.06	0.94				
		All	0.11	0.82	-0.98				
		All	0.14	0.69	-0.22				
		All	0.08	0.57	0.25				
		All	0.20	0.72	-0.16				
		All	0.15	0.99	0.08				
		All	0.07	0.61	-0.86				
		All	0.14	1.21	-0.67				
		All	0.25	1.00	-0.13				
		All	0.07	0.68	0.05				
		All	0.22	0.79	0.16				
		All	0.17	0.49	0.96				
		All	0.00	0.50	-0.19	1.49	-1.49		
		All	0.00	0.85	-0.40	1.55	-1.55		
		All	0.00	0.62	1.81	0.85	-0.85		
		All	0.00	0.64	-0.35	2.95	-0.19	-2.76	
		All	0.00	0.74	1.64	2.48	-0.08	-2.40	
		All	0.00	0.77	-0.80	1.40	0.32	-1.72	
		All	0.00	0.78	1.59	1.71	0.01	-1.72	
		All	0.00	0.67	1.40	3.54	0.73	-1.15	-3.12
		All	0.00	0.93	-0.04	2.46	0.72	-0.79	-2.39
		All	0.00	0.90	1.64	2.23	0.81	-0.72	-2.31
	▨	1	0.12	0.61	-1.55				
		1	0.19	0.98	0.88				
		1	0.10	0.60	-1.51				
		1	0.24	0.72	0.47				
		1	0.12	0.59	-1.81				
		1	0.13	0.51	0.60				

Continued, next page.

Table A.3., cont'd.:

Table A.3. IRT Generating Parameters, Administration 2, Conditions 0-16

						Parameter			
	Condition Description (*a-shift, b-shift*)	Form #	*c*	*a*	*b*	*b1*	*b2*	*b3*	*b4*
Linking Items Administration 2 No Parameter Shifts	▨	1	0.20	1.07	-0.08				
		1	0.00	1.01	0.52	1.02	0.33	-0.35	-1.00
		2	0.30	1.02	0.95				
		2	0.42	0.87	0.21				
		2	0.10	0.34	0.25				
		2	0.27	0.87	0.75				
		2	0.15	0.50	-0.83				
		2	0.20	0.48	-1.74				
		2	0.27	0.92	0.72				
		2	0.13	0.70	-0.70				
		2	0.24	0.46	1.08				
		2	0.00	1.01	0.27	2.25	0.72	-0.74	-2.23
		3	0.21	0.58	1.43				
		3	0.18	0.53	0.66				
		3	0.15	0.78	-1.21				
		3	0.19	0.82	-0.41				
		3	0.06	0.64	-0.30				
		3	0.13	0.66	-0.43				
		3	0.26	0.61	-0.10				
		3	0.22	0.79	-0.51				
		3	0.00	0.99	0.43	1.74	0.68	-0.28	-2.15
		4	0.24	0.63	-1.43				
		4	0.12	0.42	-0.31				
		4	0.39	0.50	0.22				
		4	0.11	0.79	-0.87				
		4	0.19	0.61	0.63				
		4	0.14	0.65	-0.78				
		4	0.06	0.65	-0.76				
		4	0.00	0.83	0.79	2.15	0.66	-0.81	-2.00
Linking Items Administration 2 Parameter Shifts Condition 0	No Change	1	0.14	0.99	-0.55				
		1	0.34	0.86	0.52				
		3	0.27	0.95	0.24				
		4	0.09	0.76	-0.67				
		4	0.14	0.76	1.89				
Condition 1	1 Drift Item, Hard *(-0.3, -0.5)*	1	0.14	0.99	-0.55				
		1	0.34	0.86	0.52				
		3	0.27	0.95	0.24				
		4	0.09	0.76	-0.67				
		4	*0.14*	*0.46*	*1.39*				
Condition 2	1 Drift Item, Hard *(-0.7, -0.5)*	1	0.14	0.99	-0.55				
		1	0.34	0.86	0.52				
		3	0.27	0.95	0.24				
		4	0.09	0.76	-0.67				
		4	*0.14*	*0.06*	*1.39*				

Continued, next page.

Table A.3., cont'd.:

Table A.3. IRT Generating Parameters, Administration 2, Conditions 0-16

			Parameter						
	Condition Description (*a-shift, b-shift)*	Form #	*c*	*a*	*b*	*b1*	*b2*	*b3*	*b4*
Condition 3	1 Drift Item, Hard *(-0.3, -0.8)*	1	0.14	0.99	-0.55				
		1	0.34	0.86	0.52				
		3	0.27	0.95	0.24				
		4	0.09	0.76	-0.67				
		4	*0.14*	*0.46*	*1.09*				
Condition 4	1 Drift Item, Hard *(-0.7, -0.8)*	1	0.14	0.99	-0.55				
		1	0.34	0.86	0.52				
		3	0.27	0.95	0.24				
		4	0.09	0.76	-0.67				
		4	*0.14*	*0.06*	*1.09*				
Condition 5	3 Drift Items, Spread *(-0.3, -0.5)*	1	0.14	0.99	-0.55				
		1	0.34	0.86	0.52				
		3	*0.27*	*0.65*	*-0.26*				
		4	*0.09*	*0.46*	*-1.17*				
		4	*0.14*	*0.46*	*1.39*				
Condition 6	3 Drift Items, Spread *(-0.7, -0.5)*	1	0.14	0.99	-0.55				
		1	0.34	0.86	0.52				
		3	*0.27*	*0.25*	*-0.26*				
		4	*0.09*	*0.06*	*-1.17*				
		4	*0.14*	*0.06*	*1.39*				
Condition 7	3 Drift Items, Spread *(-0.3, -0.8)*	1	0.14	0.99	-0.55				
		1	0.34	0.86	0.52				
		3	*0.27*	*0.65*	*-0.56*				
		4	*0.09*	*0.46*	*-1.47*				
		4	*0.14*	*0.46*	*1.09*				
Condition 8	3 Drift Items, Spread *(-0.7, -0.8)*	1	0.14	0.99	-0.55				
		1	0.34	0.86	0.52				
		3	*0.27*	*0.25*	*-0.56*				
		4	*0.09*	*0.06*	*-1.47*				
		4	*0.14*	*0.06*	*1.09*				
Condition 9	3 Drift Items, Moderate *(-0.3, -0.5)*	1	0.14	0.99	-0.55				
		1	*0.34*	*0.56*	*0.02*				
		3	*0.27*	*0.65*	*-0.26*				
		4	0.09	0.76	-0.67				
		4	*0.14*	*0.46*	*1.39*				
Condition 10	3 Drift Items, Moderate *(-0.7, -0.5)*	1	0.14	0.99	-0.55				
		1	*0.34*	*0.16*	*0.02*				
		3	*0.27*	*0.25*	*-0.26*				
		4	0.09	0.76	-0.67				
		4	*0.14*	*0.06*	*1.39*				
Condition 11	3 Drift Items, Moderate *(-0.3, -0.8)*	1	0.14	0.99	-0.55				
		1	*0.34*	*0.56*	*-0.28*				
		3	*0.27*	*0.65*	*-0.56*				
		4	0.09	0.76	-0.67				
		4	*0.14*	*0.46*	*1.09*				

Continued, next page.

Table A.3., cont'd.:

Table A.3. IRT Generating Parameters, Administration 2, Conditions 0-16

			Parameter						
	Condition Description (*a-shift*, *b-shift*)	Form #	*c*	*a*	*b*	*b1*	*b2*	*b3*	*b4*
Condition 12	3 Drift Items, Moderate *(-0.7, -0.8)*	1	0.14	0.99	-0.55				
		1	*0.34*	*0.16*	*-0.28*				
		3	*0.27*	*0.25*	*-0.56*				
		4	0.09	0.76	-0.67				
		4	*0.14*	*0.06*	*1.09*				
Condition 13	5 Drift Items, Spread *(-0.3, -0.5)*	*1*	*0.14*	*0.69*	*-1.05*				
		1	*0.34*	*0.56*	*0.02*				
		3	*0.27*	*0.65*	*-0.26*				
		4	*0.09*	*0.46*	*-1.17*				
		4	*0.14*	*0.46*	*1.39*				
Condition 14	5 Drift Items, Spread *(-0.7, -0.5)*	*1*	*0.14*	*0.29*	*-1.05*				
		1	*0.34*	*0.16*	*0.02*				
		3	*0.27*	*0.25*	*-0.26*				
		4	*0.09*	*0.06*	*-1.17*				
		4	*0.14*	*0.06*	*1.39*				
Condition 15	5 Drift Items, Spread *(-0.3, -0.8)*	*1*	*0.14*	*0.69*	*-1.35*				
		1	*0.34*	*0.56*	*-0.28*				
		3	*0.27*	*0.65*	*-0.56*				
		4	*0.09*	*0.46*	*-1.47*				
		4	*0.14*	*0.46*	*1.09*				
Condition 16	5 Drift Items, Spread *(-0.7, -0.8)*	*1*	*0.14*	*0.29*	*-1.35*				
		1	*0.34*	*0.16*	*-0.28*				
		3	*0.27*	*0.25*	*-0.56*				
		4	*0.09*	*0.06*	*-1.47*				
		4	*0.14*	*0.06*	*1.09*				

Table A.4. Descriptive Statistics of Examinee Ability Populations, Administrations 1 & 2

	Set						Standard		
Administration	Description	Dimension	*1*	*2*	Mean	Median	Deviation	Skewness	Kurtosis
1	1	*1*	1.00	0.80	0.00	-0.01	1.01	0.04	0.02
		2	0.80	1.00	0.01	0.00	1.00	0.05	0.02
	2	*1*	1.00	0.60	0.00	-0.01	1.01	0.04	0.02
		2	0.60	1.00	0.00	-0.01	1.00	0.05	0.01
2	No Shift	▨	▨	▨	0.01	-0.01	1.01	-0.02	0.08
	Shift	▨	▨	▨	0.26	0.14	0.89	0.17	0.50

Table A.5. Power & Type I Error for the Delta Plot Method of Item Parameter Drift Detection Using Three Standard Deviations as the Criterion, by Condition

	Condition	Amount of IPD		Power					Type I
Distribution	Description	*a*	*b*	Item 1	Item 2	Item 3	Item 4	Item 5	Error
Normal (No Shift)	No IPD	None	None	▨	▨	▨	▨	▨	0.000
	1 Item, Hard	-0.3	-0.5	0.970					0.000
		-0.7	-0.5	1.000					0.000
		-0.3	-0.8	1.000					0.000
		-0.7	-0.8	1.000					0.000
	3 Items, Spread	-0.3	-0.5	0.740	0.000	0.000			0.000
		-0.7	-0.5	1.000	0.000	0.000			0.000
		-0.3	-0.8	1.000	0.000	0.000			0.000
		-0.7	-0.8	1.000	0.000	0.000			0.000
	3 Items, Moderate	-0.3	-0.5	1.000		0.000	0.000		0.000
		-0.7	-0.5	1.000		0.000	0.000		0.000
		-0.3	-0.8	1.000		0.000	0.000		0.000
		-0.7	-0.8	1.000		0.000	0.000		0.000
	5 Items, Spread	-0.3	-0.5	1.000	0.000	0.000	0.000	0.000	0.000
		-0.7	-0.5	1.000	0.000	0.000	0.000	0.000	0.000
		-0.3	-0.8	1.000	0.000	0.000	0.000	0.000	0.000
		-0.7	-0.8	1.000	0.000	0.000	0.000	0.000	0.000
Negatively Skewed (Shift)	No IPD	None	None	▨	▨	▨	▨	▨	0.000
	1 Item, Hard	-0.3	-0.5	0.000					0.000
		-0.7	-0.5	0.010					0.000
		-0.3	-0.8	1.000					0.000
		-0.7	-0.8	0.990					0.000
	3 Items, Spread	-0.3	-0.5	0.000	0.000	0.000			0.000
		-0.7	-0.5	0.000	0.000	0.000			0.000
		-0.3	-0.8	0.670	0.000	0.000			0.000
		-0.7	-0.8	0.470	0.000	0.000			0.000
	3 Items, Moderate	-0.3	-0.5	0.000		0.000	0.000		0.000
		-0.7	-0.5	1.000		0.000	0.000		0.000
		-0.3	-0.8	0.840		0.000	0.000		0.000
		-0.7	-0.8	1.000		0.000	0.000		0.000
	5 Items, Spread	-0.3	-0.5	0.000	0.000	0.000	0.000	0.000	0.000
		-0.7	-0.5	1.000	0.000	0.000	0.000	0.000	0.000
		-0.3	-0.8	0.000	0.000	0.000	0.000	0.000	0.000
		-0.7	-0.8	1.000	0.000	0.000	0.000	0.000	0.000

Table A.6. Power & Type I Error for the *b* Plot Method of Item Parameter Drift Detection Using Three Standard Deviations as the Criterion, by Condition

	Condition	Amount of IPD		Power					Type I
Distribution	Description	*a*	*b*	Item 1	Item 2	Item 3	Item 4	Item 5	Error
Normal (No Shift)	No IPD	None	None	▨	▨	▨	▨	▨	0.015
	1 Item, Hard	-0.3	-0.5	0.730					0.007
		-0.7	-0.5	1.000					0.000
		-0.3	-0.8	0.990					0.004
		-0.7	-0.8	1.000					0.000
	3 Items, Spread	-0.3	-0.5	0.690	0.000	0.000			0.006
		-0.7	-0.5	0.690	0.720	0.000			0.001
		-0.3	-0.8	0.910	0.000	0.000			0.002
		-0.7	-0.8	0.630	0.800	0.000			0.001
	3 Items, Moderate	-0.3	-0.5	0.580		0.000	0.000		0.024
		-0.7	-0.5	0.900		0.000	0.150		0.001
		-0.3	-0.8	0.870		0.000	0.270		0.002
		-0.7	-0.8	0.860		0.000	0.130		0.002
	5 Items, Spread	-0.3	-0.5	0.840	0.000	0.000	0.190	0.000	0.001
		-0.7	-0.5	0.650	0.720	0.000	0.030	0.000	0.001
		-0.3	-0.8	0.630	0.760	0.000	0.130	0.000	0.000
		-0.7	-0.8	0.470	0.800	0.000	0.110	0.000	0.000
Negatively Skewed (Shift)	No IPD	None	None	▨	▨	▨	▨	▨	0.018
	1 Item, Hard	-0.3	-0.5	1.000					0.001
		-0.7	-0.5	0.720					0.005
		-0.3	-0.8	1.000					0.001
		-0.7	-0.8	0.950					0.002
	3 Items, Spread	-0.3	-0.5	0.980	0.000	0.000			0.001
		-0.7	-0.5	0.710	0.000	0.000			0.004
		-0.3	-0.8	1.000	0.000	0.020			0.000
		-0.7	-0.8	0.880	0.000	0.000			0.002
	3 Items, Moderate	-0.3	-0.5	0.970		0.000	0.000		0.001
		-0.7	-0.5	0.930		0.000	0.330		0.000
		-0.3	-0.8	1.000		0.020	0.030		0.000
		-0.7	-0.8	0.950		0.000	0.200		0.001
	5 Items, Spread	-0.3	-0.5	0.950	0.000	0.000	0.000	0.000	0.001
		-0.7	-0.5	0.680	0.710	0.000	0.130	0.000	0.000
		-0.3	-0.8	0.980	0.000	0.010	0.000	0.000	0.000
		-0.7	-0.8	0.460	0.850	0.000	0.090	0.000	0.000

Table A.7. Power & Type I Error for the RPU Method of Item Parameter Drift Detection Using Conservative Criteria, by Condition

	Condition	Amount of IPD		Power					Type I
Distribution	Description	*a*	*b*	Item 1	Item 2	Item 3	Item 4	Item 5	Error
Normal (No Shift)	No IPD	None	None	▨	▨	▨	▨	▨	0.000
	1 Item, Hard	-0.3	-0.5	0.000					0.000
		-0.7	-0.5	1.000					0.000
		-0.3	-0.8	0.000					0.000
		-0.7	-0.8	1.000					0.000
	3 Items, Spread	-0.3	-0.5	0.000	0.000	0.000			0.000
		-0.7	-0.5	1.000	0.000	0.000			0.000
		-0.3	-0.8	0.000	0.060	0.130			0.000
		-0.7	-0.8	1.000	0.000	0.000			0.000
	3 Items, Moderate	-0.3	-0.5	1.000		0.000	0.000		0.000
		-0.7	-0.5	1.000		0.000	0.000		0.000
		-0.3	-0.8	1.000		0.000	0.000		0.000
		-0.7	-0.8	1.000		0.000	0.000		0.000
	5 Items, Spread	-0.3	-0.5	1.000	0.000	0.000	0.000	0.000	0.000
		-0.7	-0.5	1.000	0.000	0.000	0.000	0.000	0.000
		-0.3	-0.8	1.000	0.000	0.000	0.000	0.000	0.000
		-0.7	-0.8	1.000	0.000	0.000	0.070	0.000	0.000
Negatively Skewed (Shift)	No IPD	None	None	▨	▨	▨	▨	▨	0.000
	1 Item, Hard	-0.3	-0.5	0.000					0.000
		-0.7	-0.5	0.000					0.003
		-0.3	-0.8	0.000					0.000
		-0.7	-0.8	0.000					0.003
	3 Items, Spread	-0.3	-0.5	0.000	0.670	0.000			0.000
		-0.7	-0.5	0.000	0.630	0.010			0.003
		-0.3	-0.8	0.000	1.000	1.000			0.000
		-0.7	-0.8	0.000	1.000	1.000			0.003
	3 Items, Moderate	-0.3	-0.5	0.000		0.000	0.000		0.000
		-0.7	-0.5	1.000		0.000	0.000		0.000
		-0.3	-0.8	0.000		1.000	0.020		0.000
		-0.7	-0.8	1.000		0.000	0.000		0.000
	5 Items, Spread	-0.3	-0.5	0.000	0.630	0.000	0.000	1.000	0.000
		-0.7	-0.5	1.000	0.060	0.000	0.000	0.000	0.000
		-0.3	-0.8	0.000	1.000	1.000	0.020	1.000	0.000
		-0.7	-0.8	1.000	0.040	0.000	1.000	0.000	0.028

Table A.8. Number & Percent of Items that Required *c*-Parameter to be Set to Zero During Calibration

	Common	% Common	Equating	% Equating
Grade 4	4	10.0	1	3.0
Grade 8	1	2.0	0	0.0
Grade 11	4	10.0	1	3.0

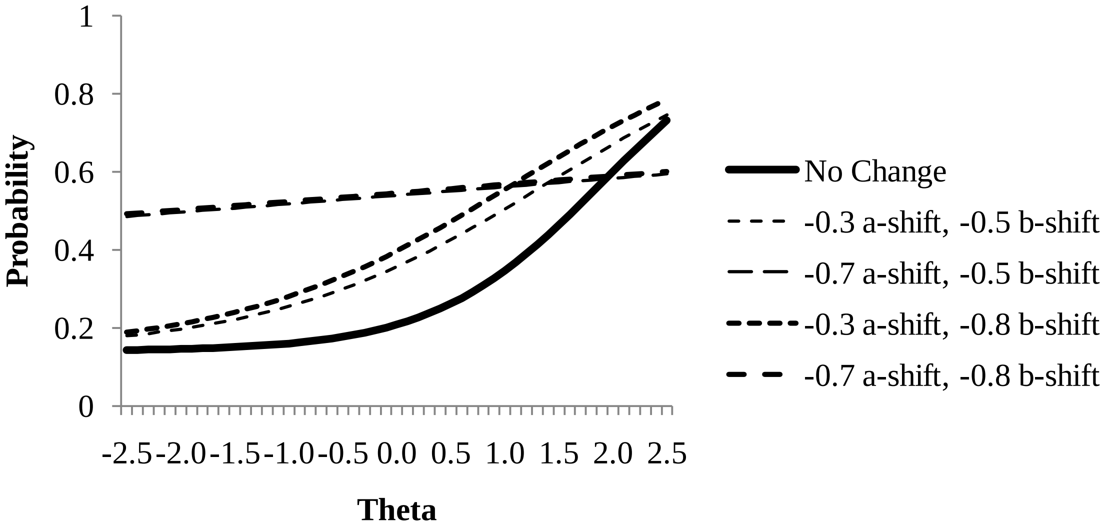


Figure A.1. Item 1, Hard

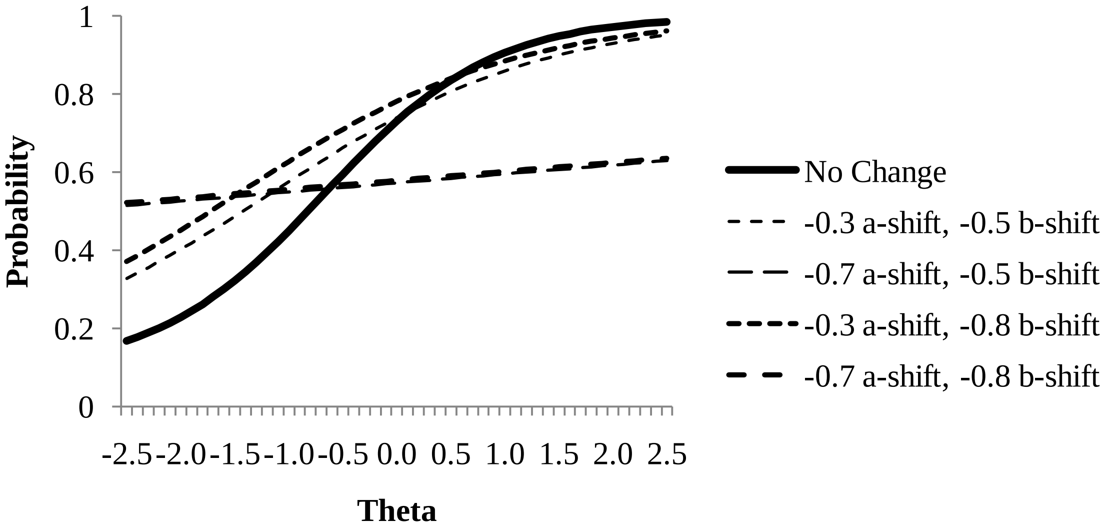


Figure A.2. Item 2, Easy, 1

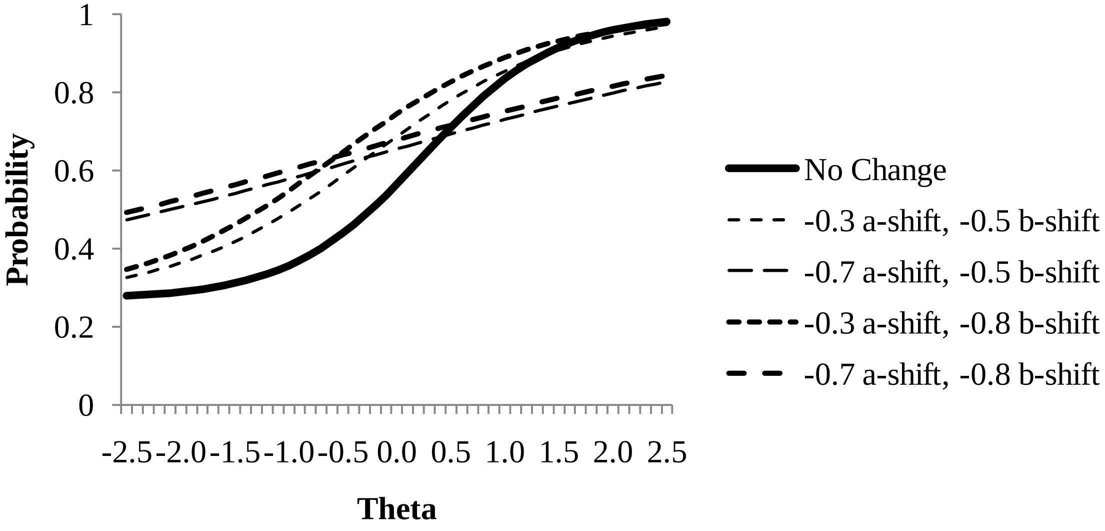


Figure A.3. Item 3, Moderate, 1

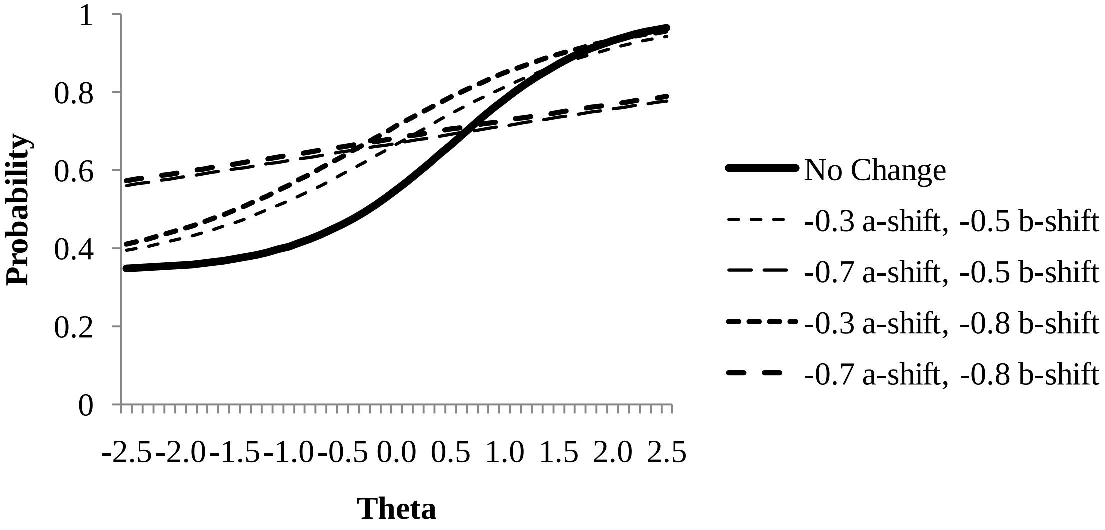


Figure A.4. Item 4, Moderate, 2

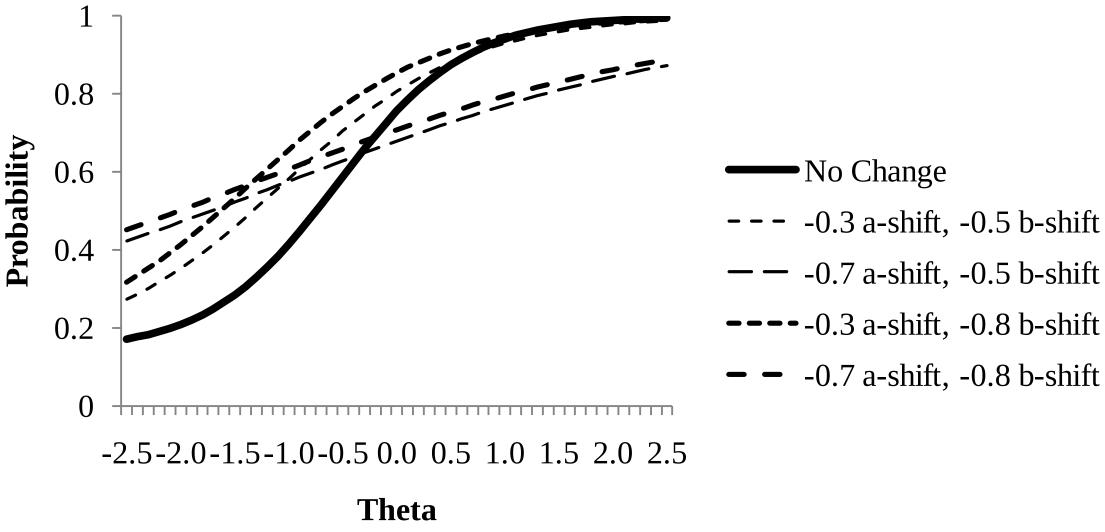
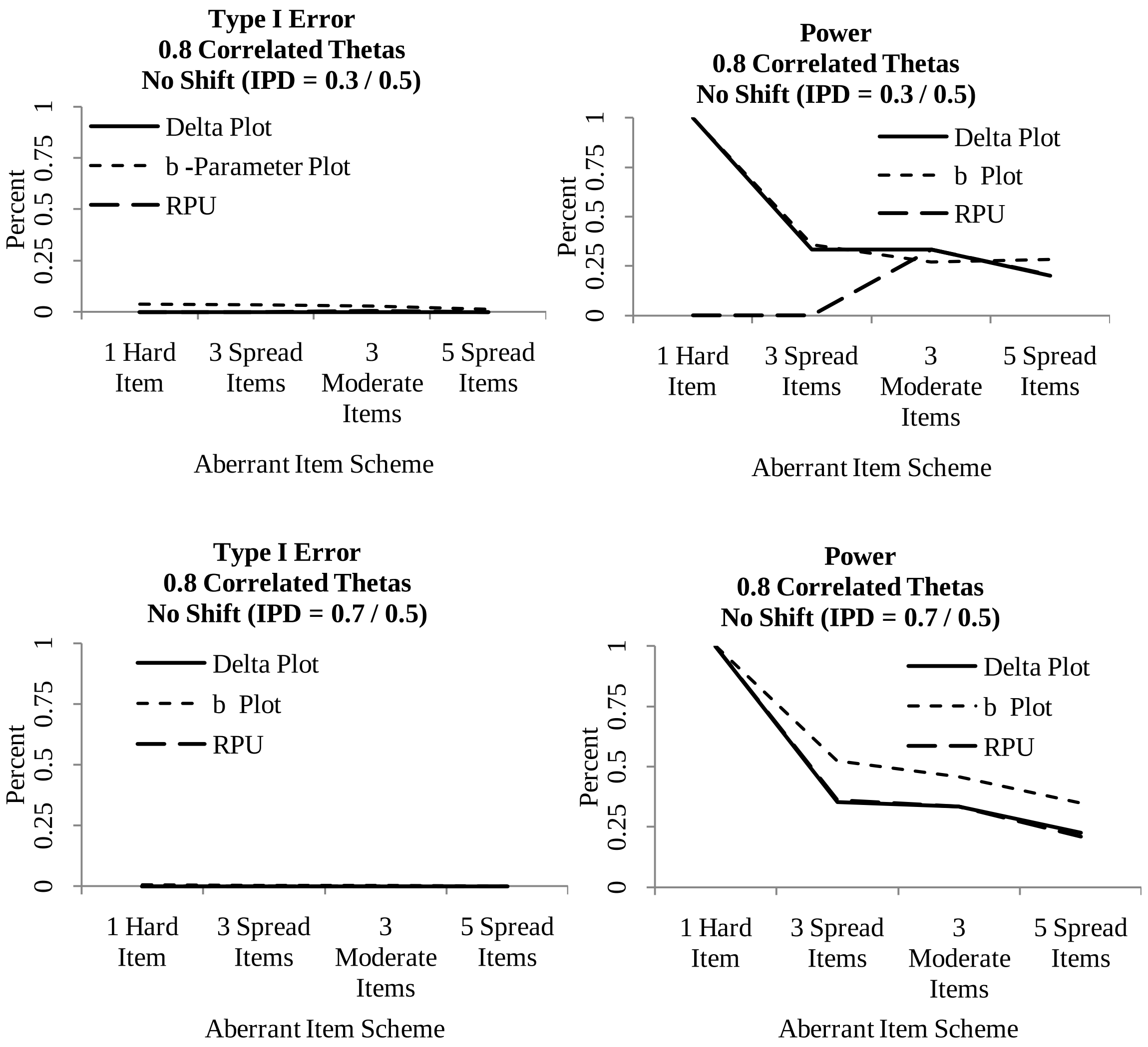


Figure A.5. Item 5, Easy, 2

APPENDIX B

TYPE I ERROR & POWER SUMMARY PLOTS



Type I Error 0.8 Correlated Thetas No Shift (IPD = 0.3 / 0.5)

Power 0.8 Correlated Thetas No Shift (IPD = 0.3 / 0.5)

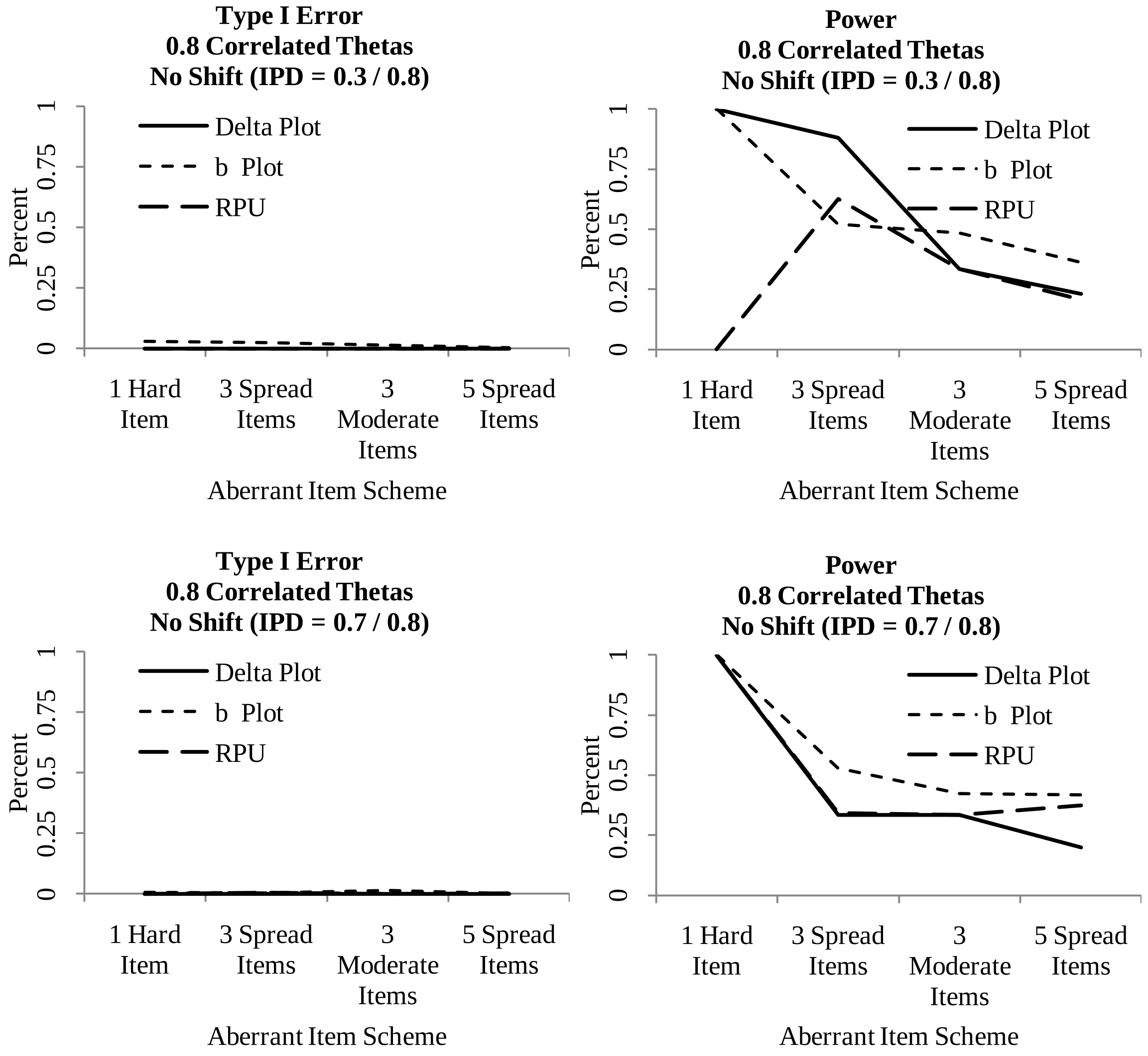
Type I Error 0.8 Correlated Thetas No Shift (IPD = 0.7 / 0.5)

Power 0.8 Correlated Thetas No Shift (IPD = 0.7 / 0.5)

Figure B.1. Type I Error & Power Summary Plots

Continued, next page.

Figure B.1., cont’d.:



Type I Error
0.8 Correlated Thetas
No Shift (IPD = 0.3 / 0.8)

Power
0.8 Correlated Thetas
No Shift (IPD = 0.3 / 0.8)

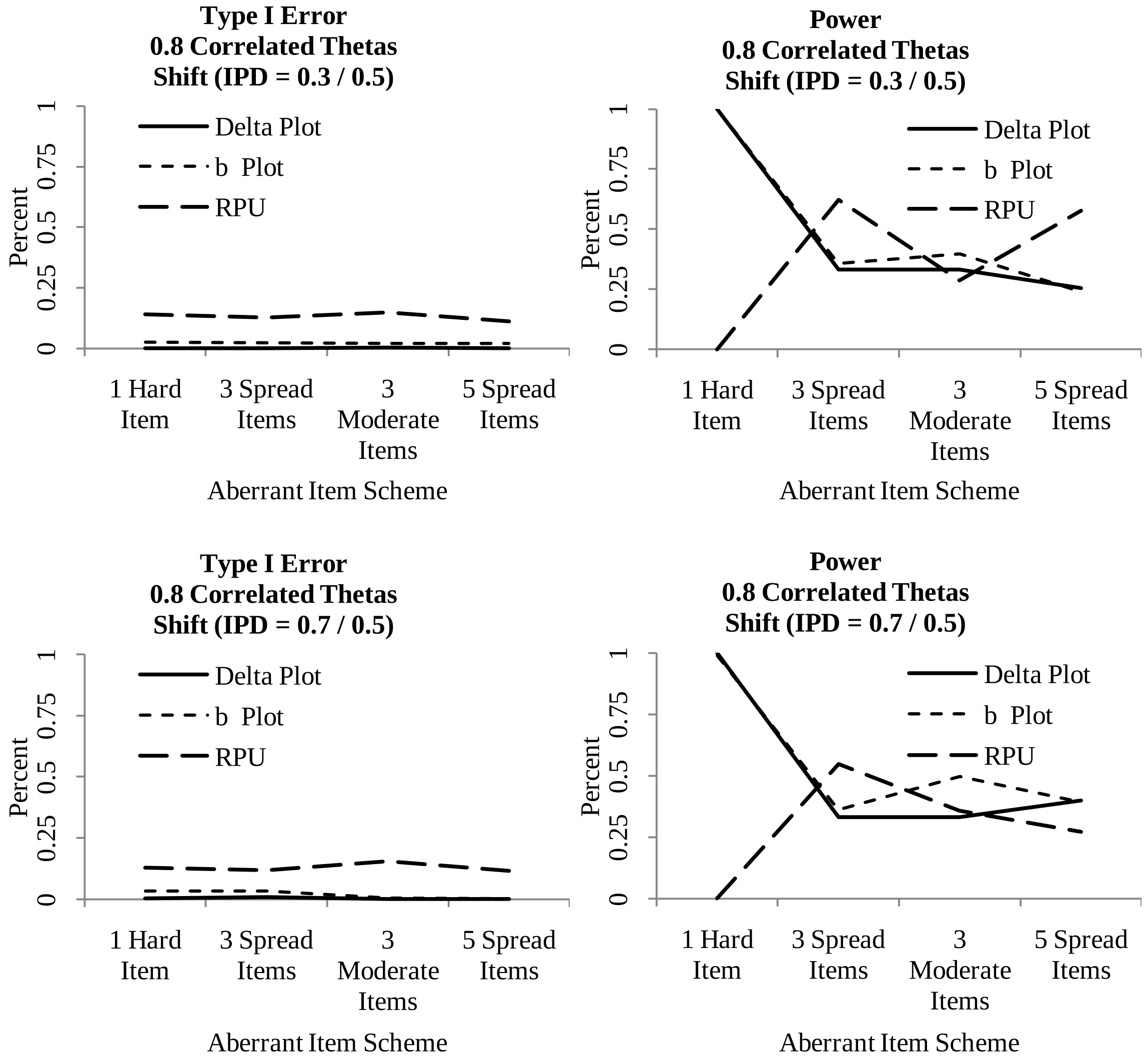
Type I Error
0.8 Correlated Thetas
No Shift (IPD = 0.7 / 0.8)

Power
0.8 Correlated Thetas
No Shift (IPD = 0.7 / 0.8)

Figure B.1. Type I Error & Power Summary Plots

Continued, next page.

Figure B.1., cont'd.:



Type I Error
0.8 Correlated Thetas
Shift (IPD = 0.3 / 0.5)

Power
0.8 Correlated Thetas
Shift (IPD = 0.3 / 0.5)

Type I Error
0.8 Correlated Thetas
Shift (IPD = 0.7 / 0.5)

Power
0.8 Correlated Thetas
Shift (IPD = 0.7 / 0.5)

Figure B.1. Type I Error & Power Summary Plots

Continued, next page.

Figure B.1., cont'd.:

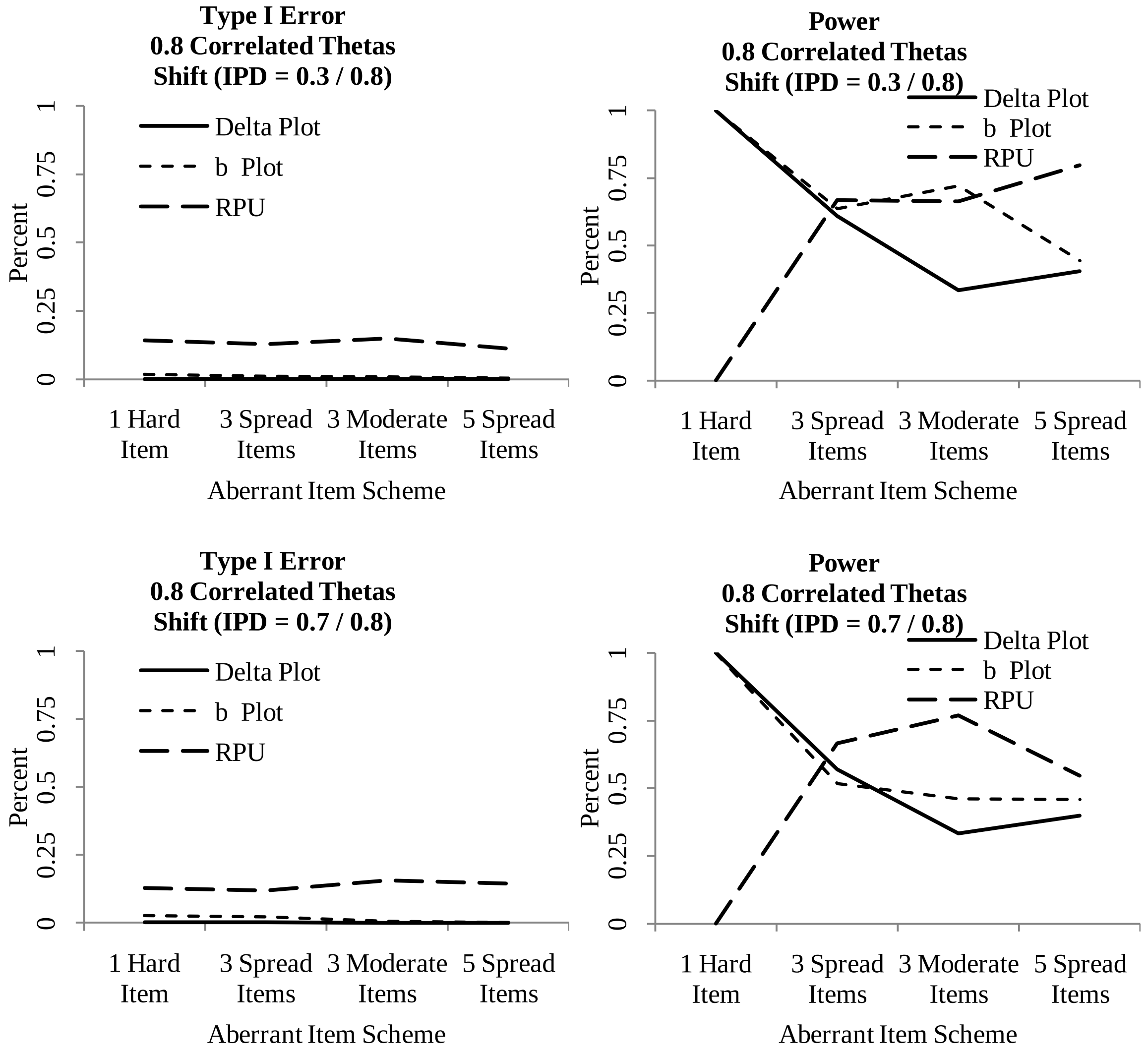
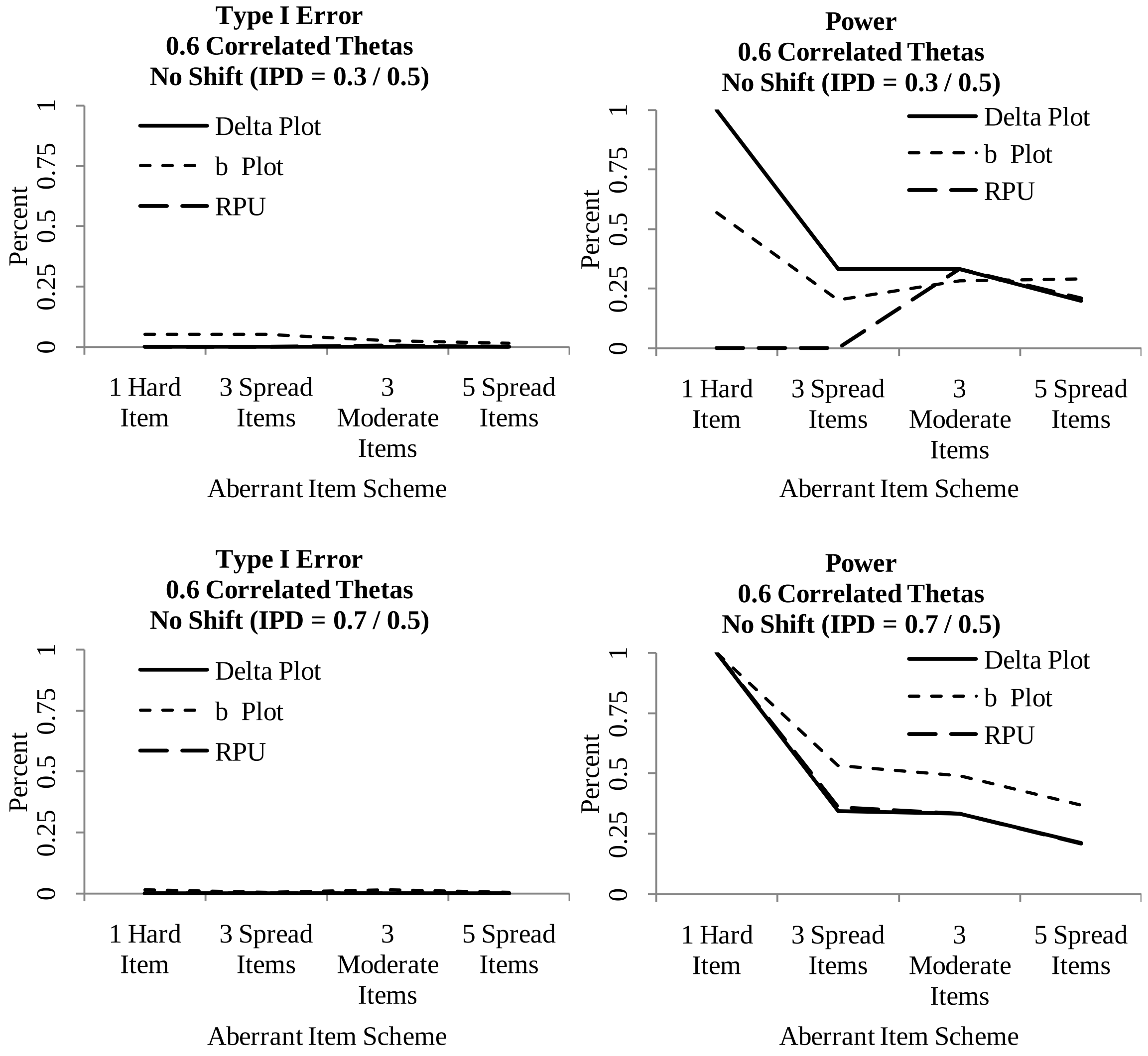


Figure B.1. Type I Error & Power Summary Plots

Continued, next page.

Figure B.1., cont'd.:


Type I Error
0.6 Correlated Thetas
No Shift (IPD = 0.3 / 0.5)

Power
0.6 Correlated Thetas
No Shift (IPD = 0.3 / 0.5)

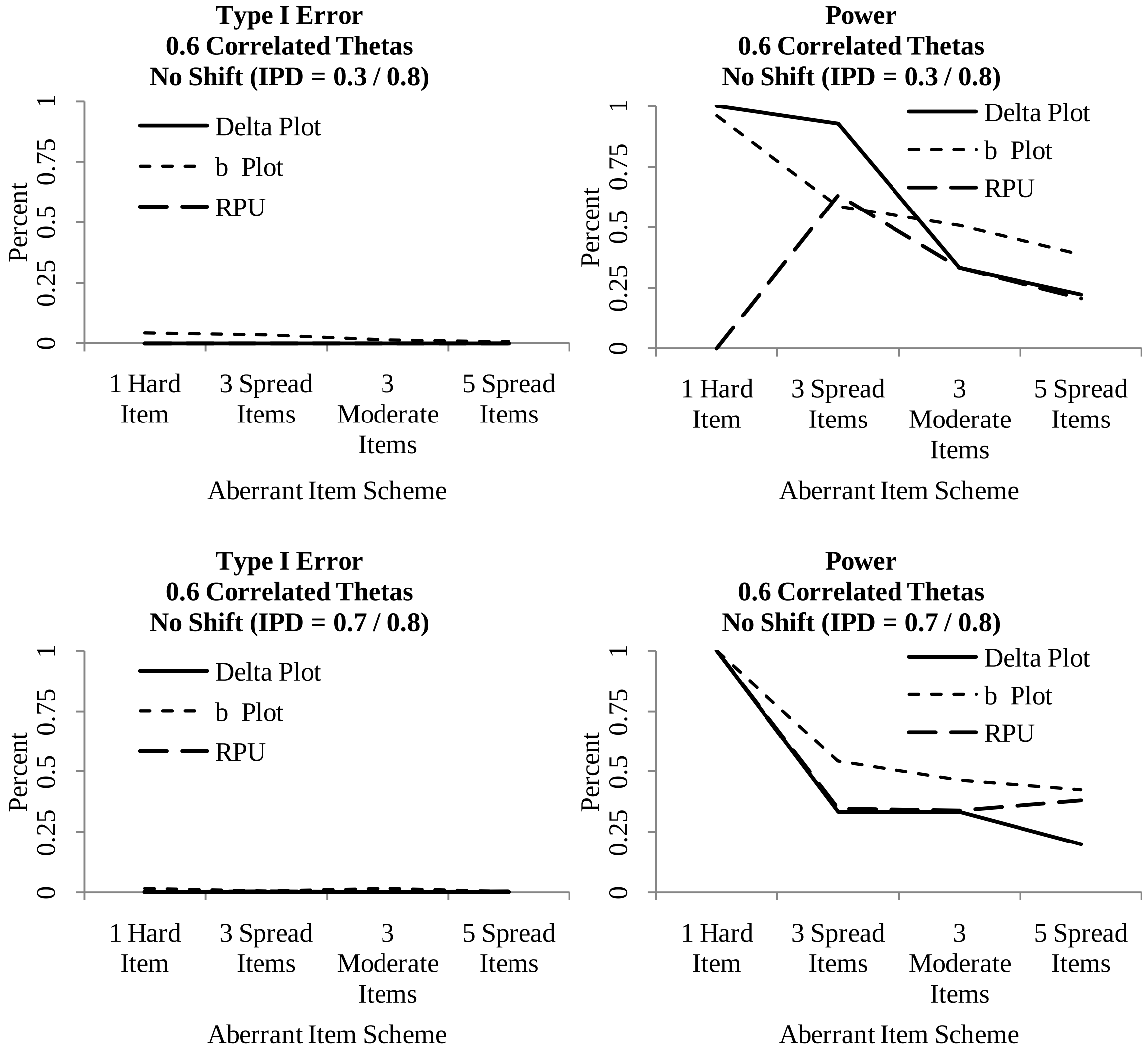
Type I Error
0.6 Correlated Thetas
No Shift (IPD = 0.7 / 0.5)

Power
0.6 Correlated Thetas
No Shift (IPD = 0.7 / 0.5)

Figure B.1. Type I Error & Power Summary Plots

Continued, next page.

Figure B.1., cont'd.:


Type I Error
0.6 Correlated Thetas
No Shift (IPD = 0.3 / 0.8)

Power
0.6 Correlated Thetas
No Shift (IPD = 0.3 / 0.8)

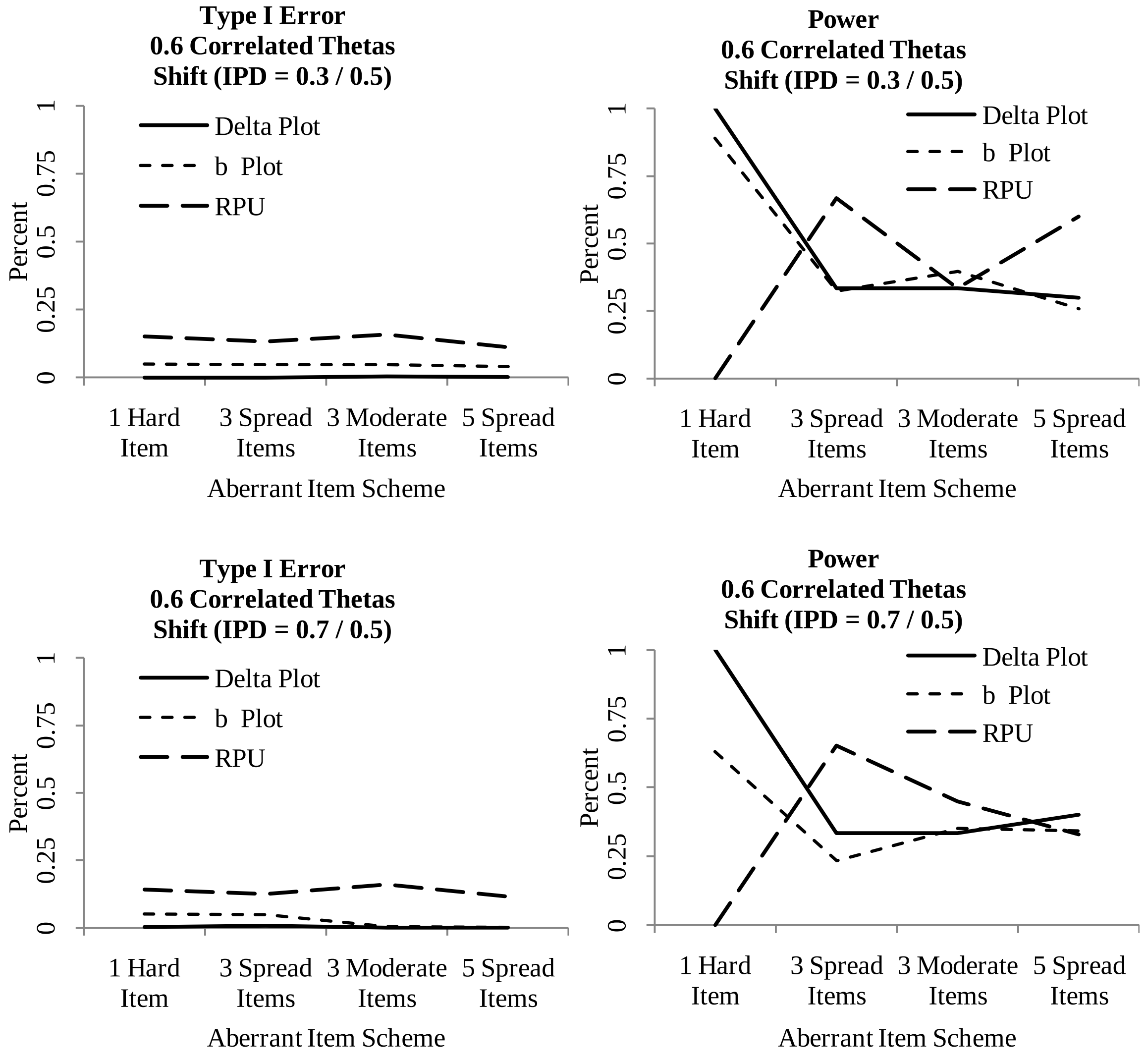
Type I Error
0.6 Correlated Thetas
No Shift (IPD = 0.7 / 0.8)

Power
0.6 Correlated Thetas
No Shift (IPD = 0.7 / 0.8)

Figure B.1. Type I Error & Power Summary Plots

Continued, next page.

Figure B.1., cont'd.:



Type I Error
0.6 Correlated Thetas
Shift (IPD = 0.3 / 0.5)

Power
0.6 Correlated Thetas
Shift (IPD = 0.3 / 0.5)

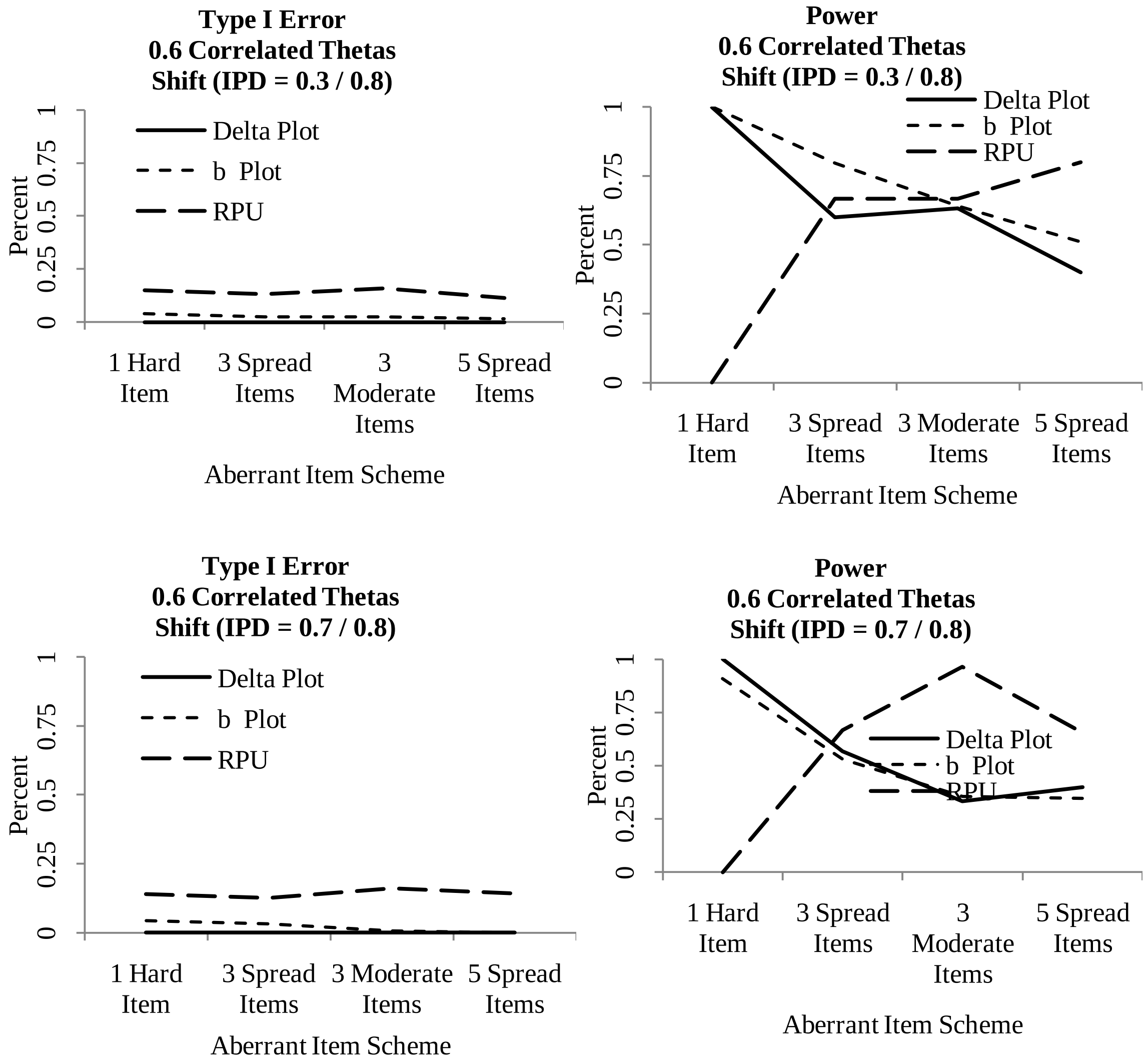
Type I Error
0.6 Correlated Thetas
Shift (IPD = 0.7 / 0.5)

Power
0.6 Correlated Thetas
Shift (IPD = 0.7 / 0.5)

Figure B.1. Type I Error & Power Summary Plots

Continued, next page.

Figure B.1., cont'd.:



Type I Error
0.6 Correlated Thetas
Shift (IPD = 0.3 / 0.8)

Power
0.6 Correlated Thetas
Shift (IPD = 0.3 / 0.8)

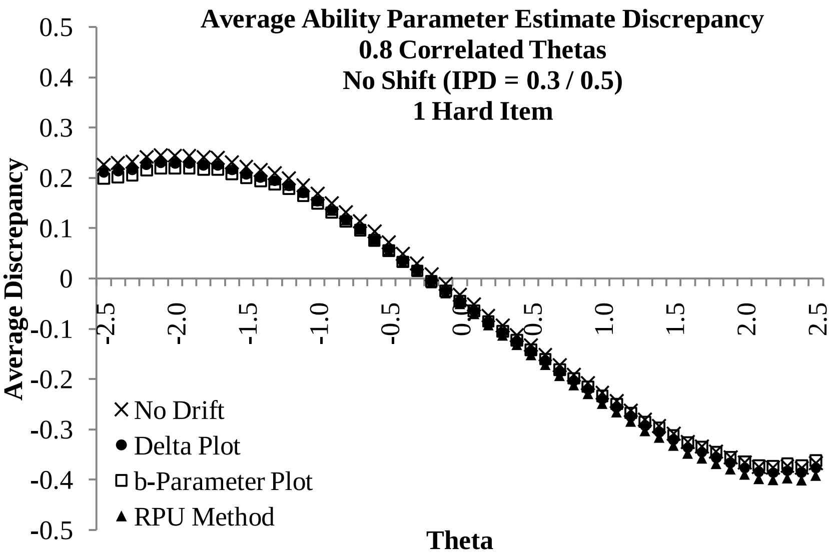
Type I Error
0.6 Correlated Thetas
Shift (IPD = 0.7 / 0.8)

Power
0.6 Correlated Thetas
Shift (IPD = 0.7 / 0.8)

Figure B.1. Type I Error & Power Summary Plots

APPENDIX C

AVERAGE ABILITY PARAMETER ESTIMATE DISCREPANCY PLOTS



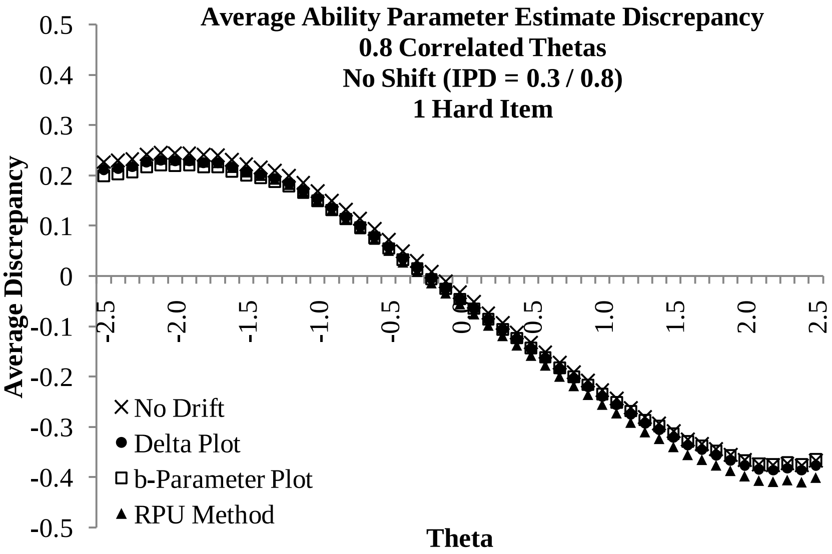
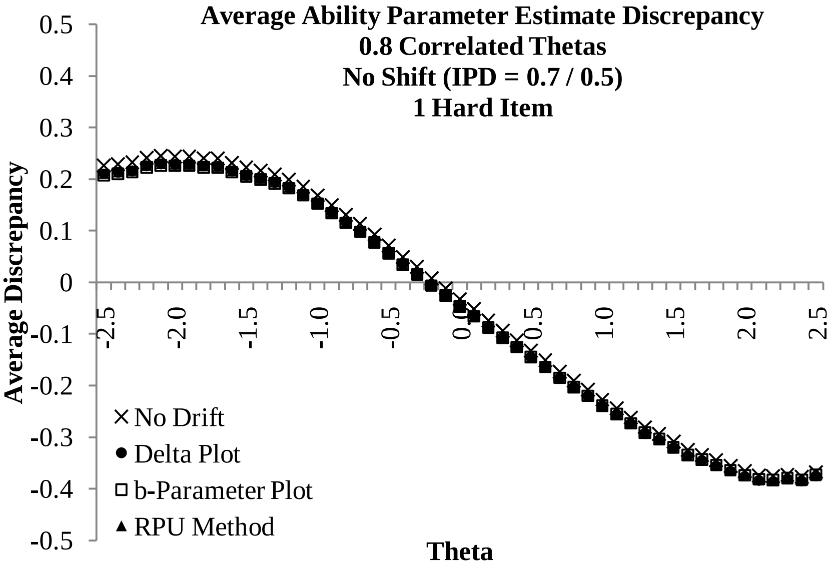


Figure C.1. Average Ability Parameter Estimate Discrepancy Plots

Continued, next page.

Figure C.1., cont'd.:



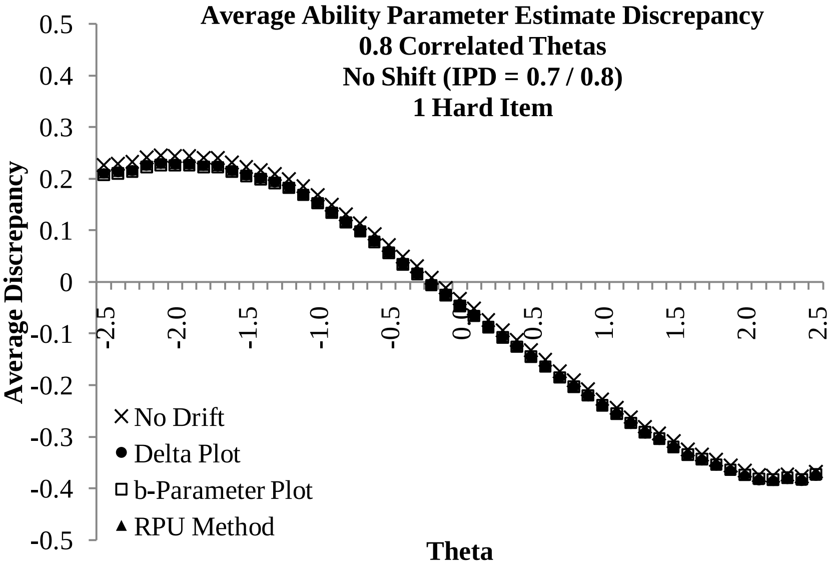
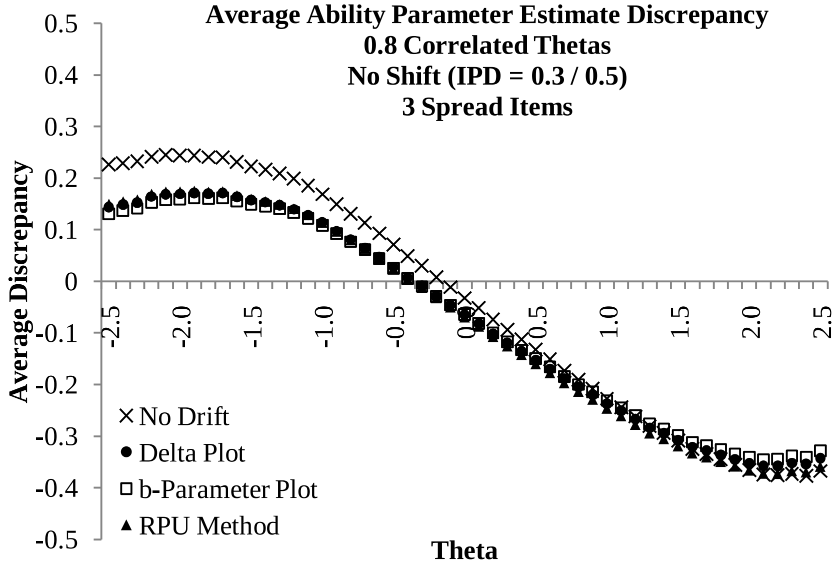


Figure C.1. Average Ability Parameter Estimate Discrepancy Plots

Continued, next page.

Figure C.1., cont'd.:



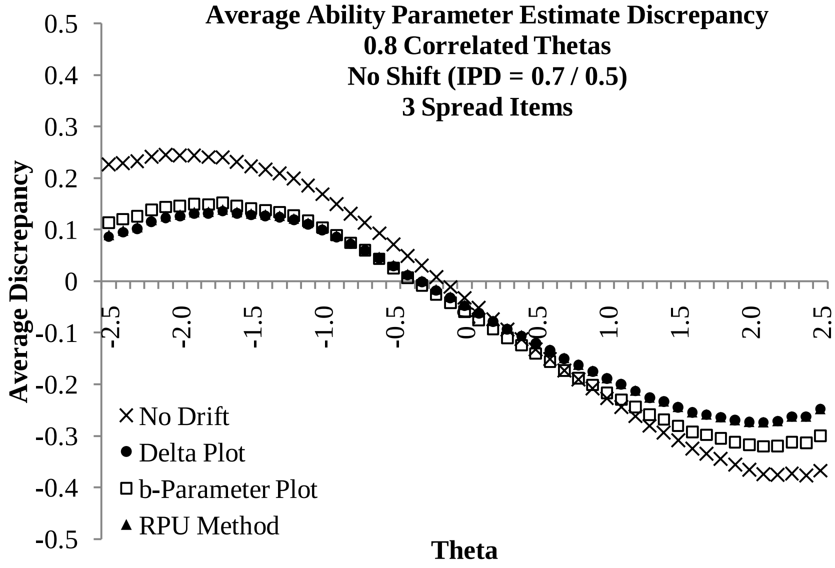
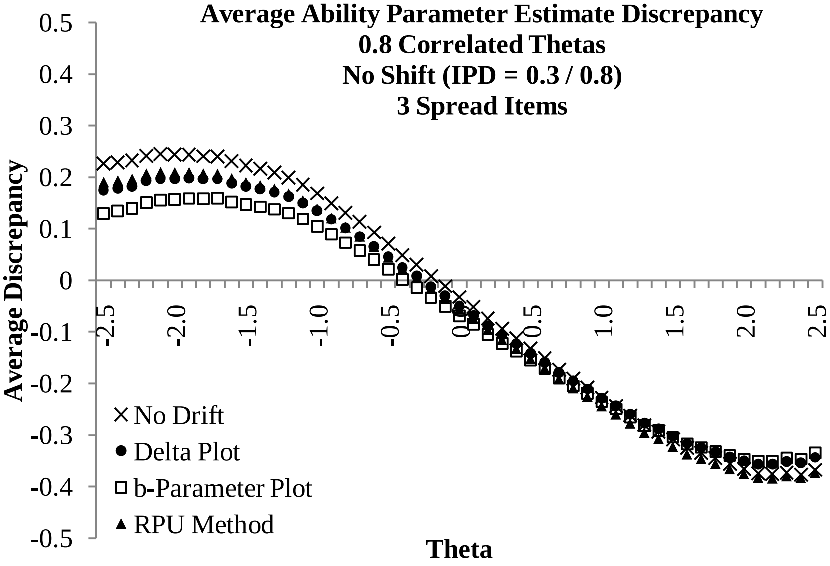


Figure C.1. Average Ability Parameter Estimate Discrepancy Plots

Continued, next page.

Figure C.1., cont'd.:



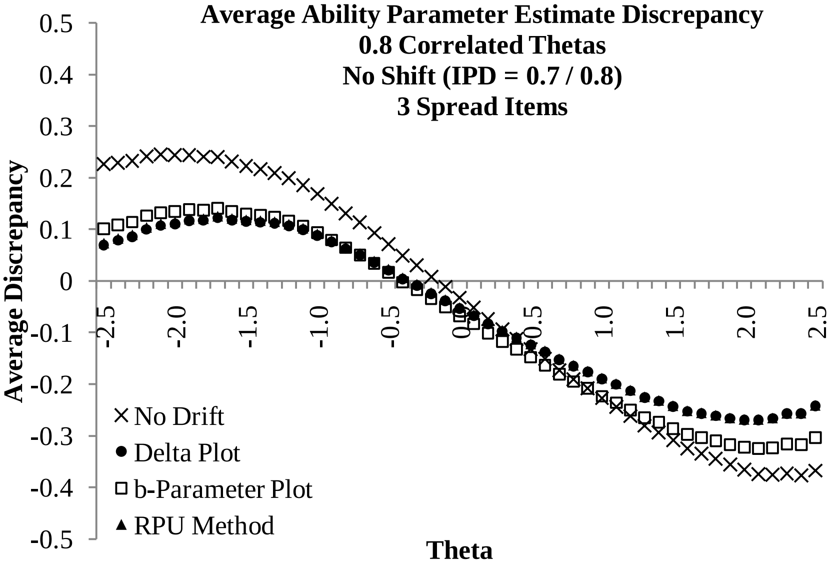
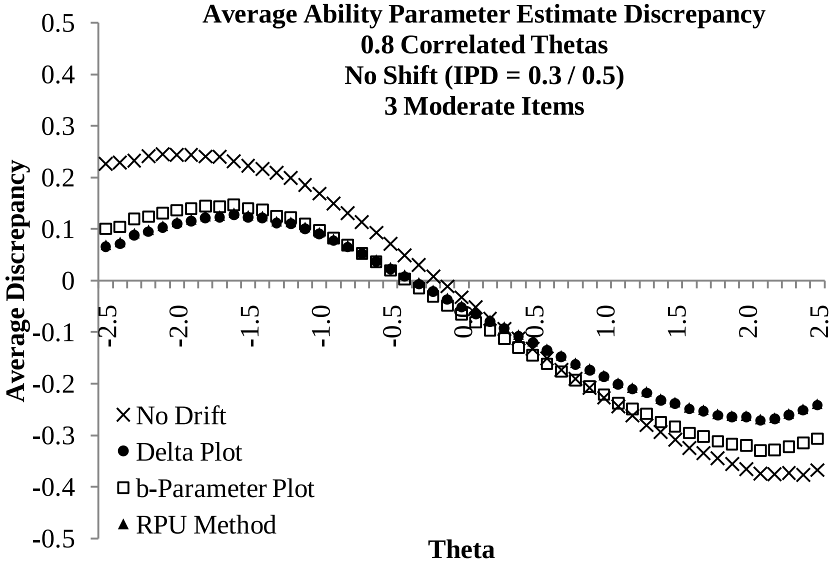


Figure C.1. Average Ability Parameter Estimate Discrepancy Plots

Continued, next page.

Figure C.1., cont'd.:



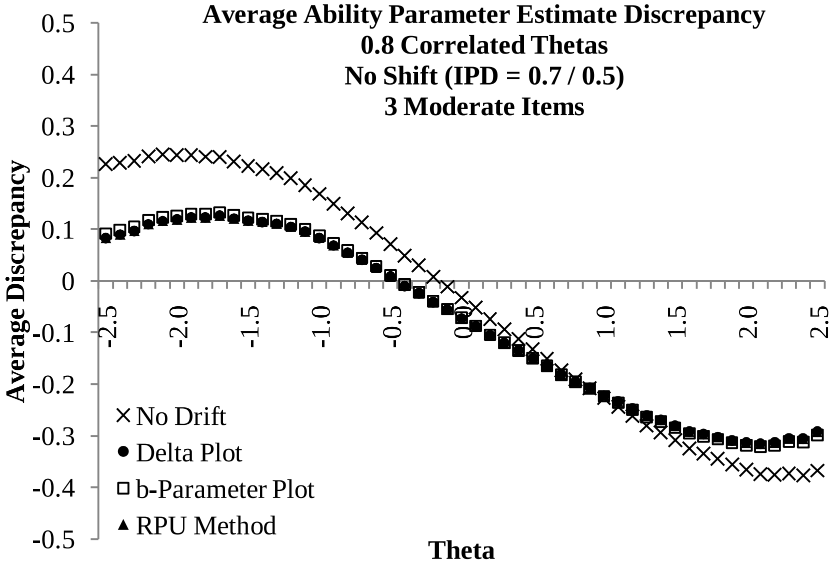
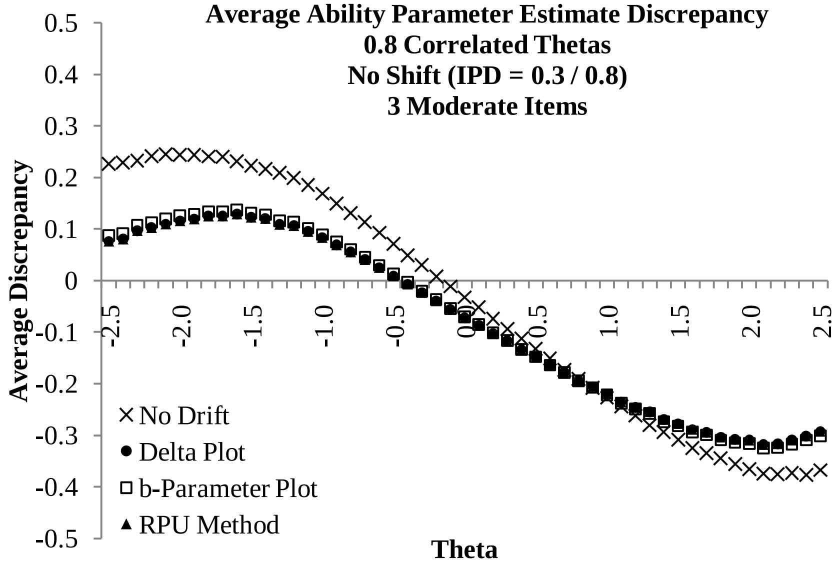


Figure C.1. Average Ability Parameter Estimate Discrepancy Plots

Continued, next page.

Figure C.1., cont'd.:



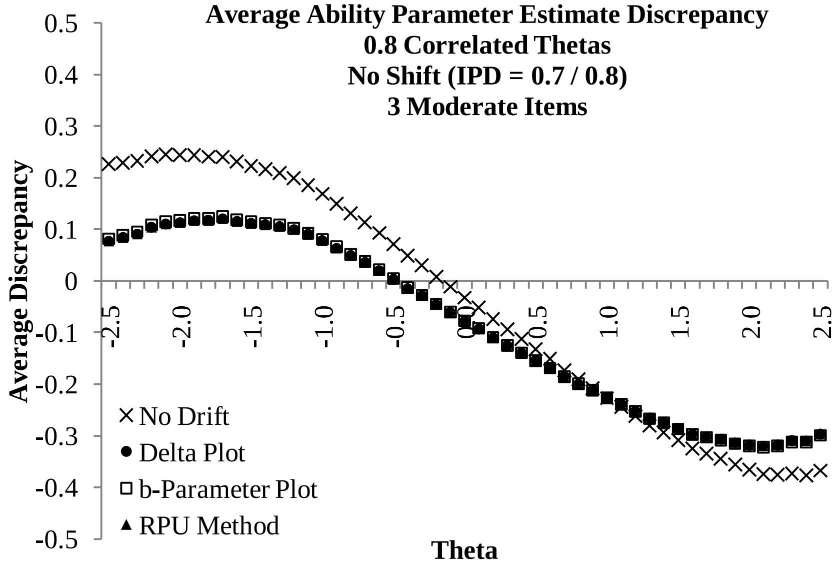
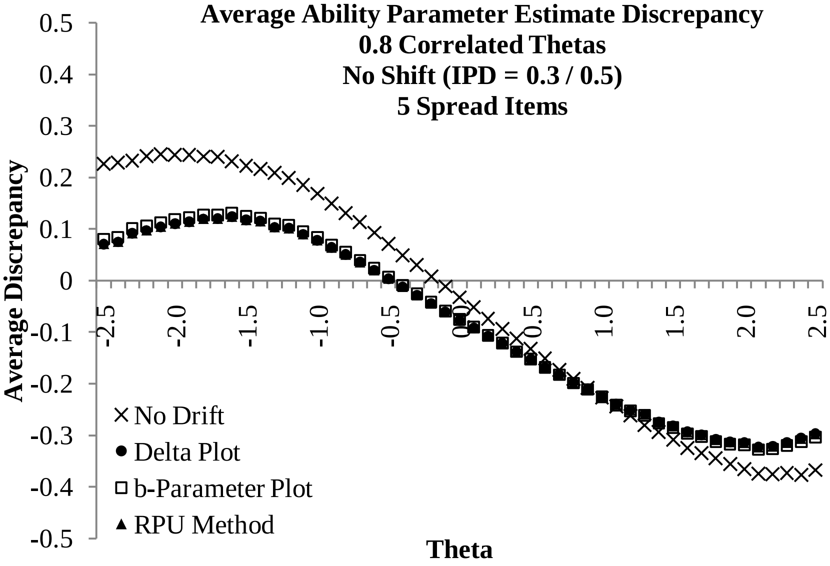


Figure C.1. Average Ability Parameter Estimate Discrepancy Plots

Continued, next page.

Figure C.1., cont'd.:



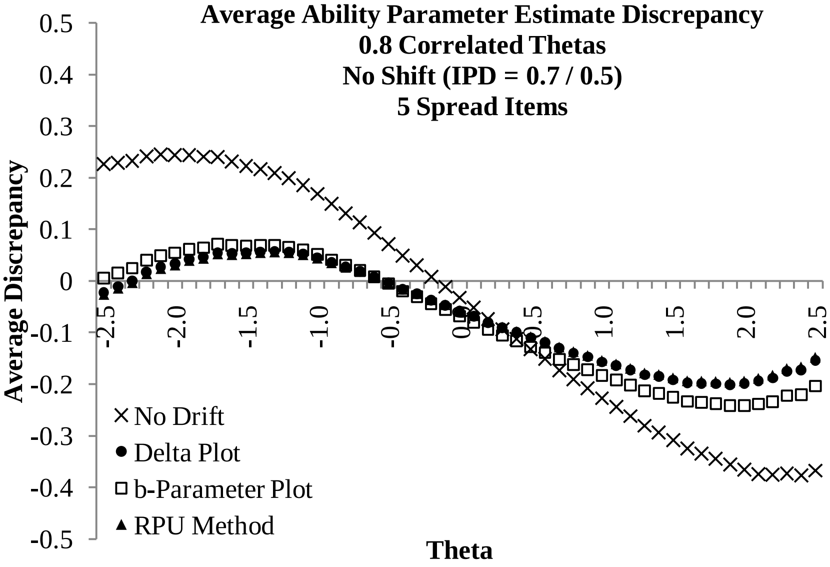
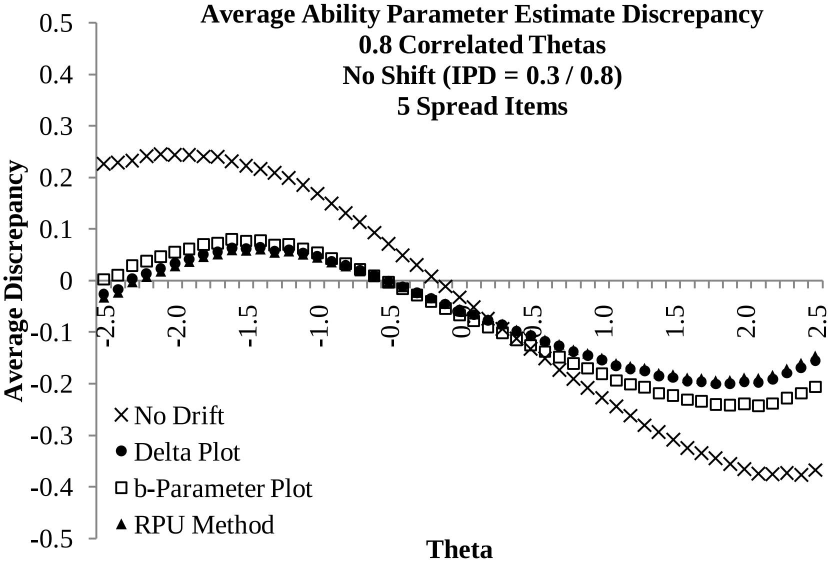


Figure C.1. Average Ability Parameter Estimate Discrepancy Plots

Continued, next page.

Figure C.1., cont'd.:



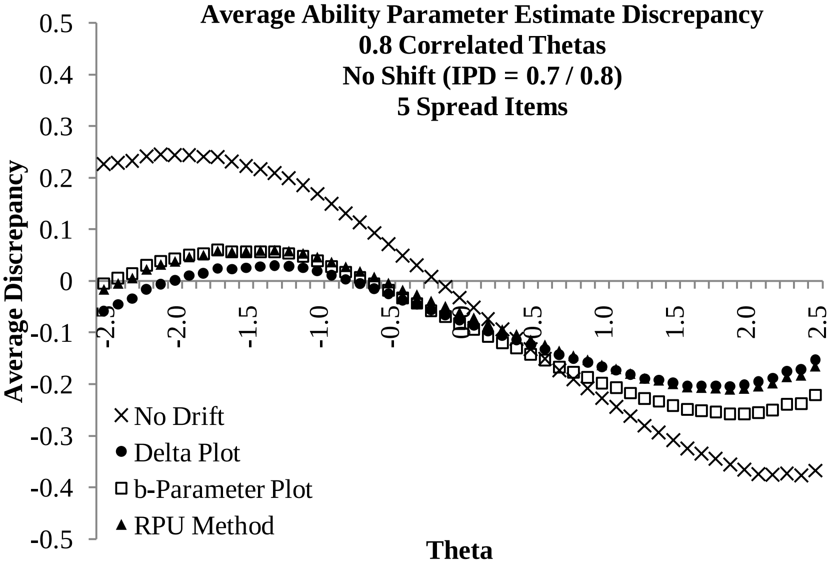
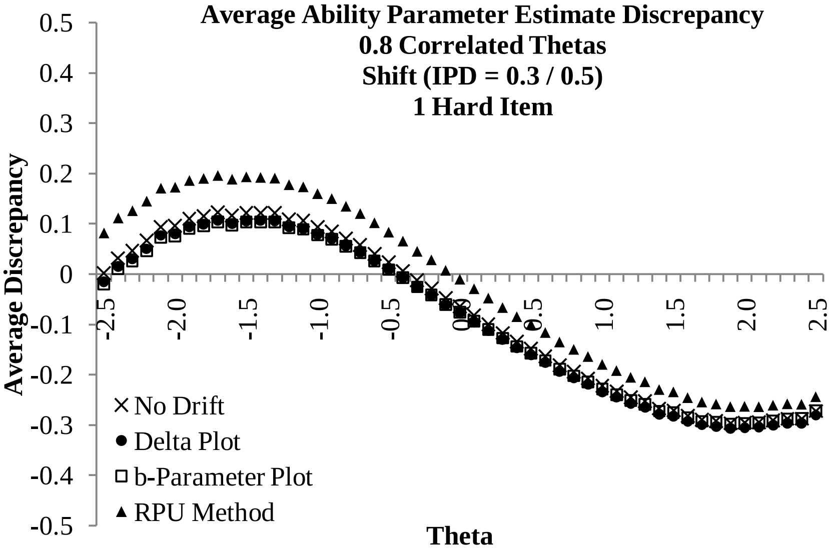


Figure C.1. Average Ability Parameter Estimate Discrepancy Plots

Continued, next page.

Figure C.1., cont'd.:



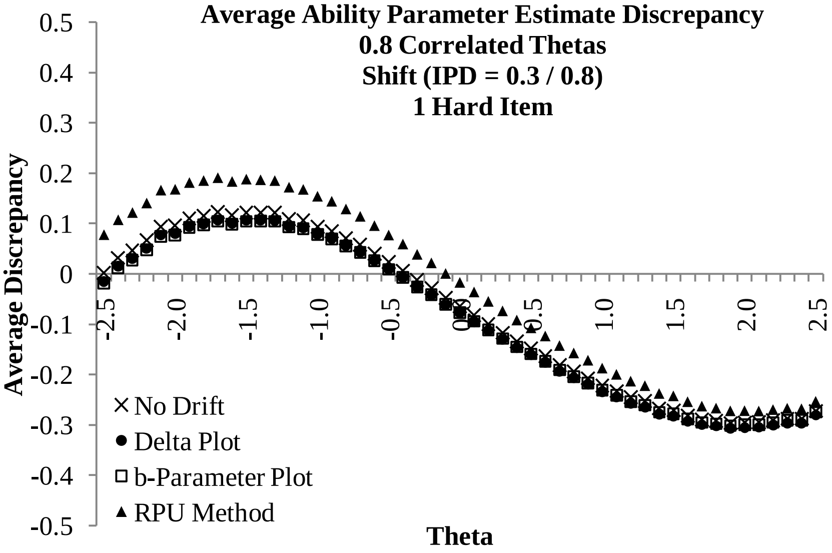
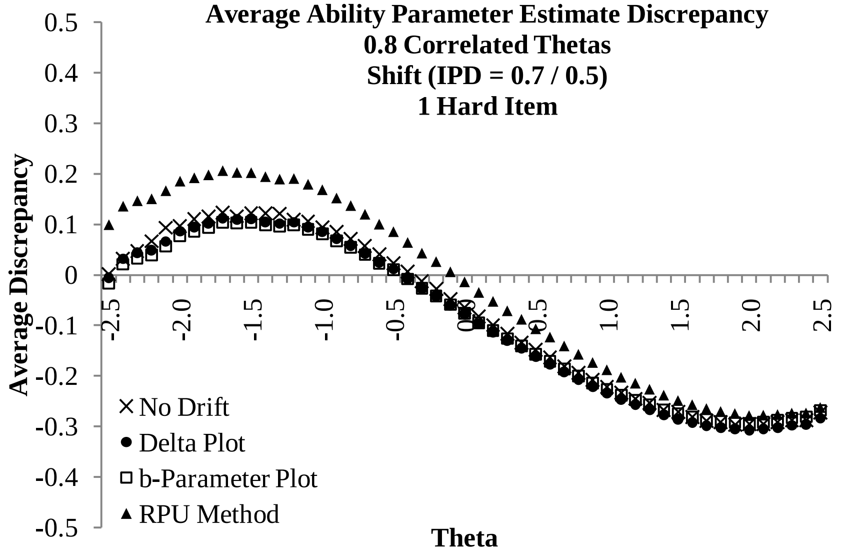


Figure C.1. Average Ability Parameter Estimate Discrepancy Plots

Continued, next page.

Figure C.1., cont'd.:



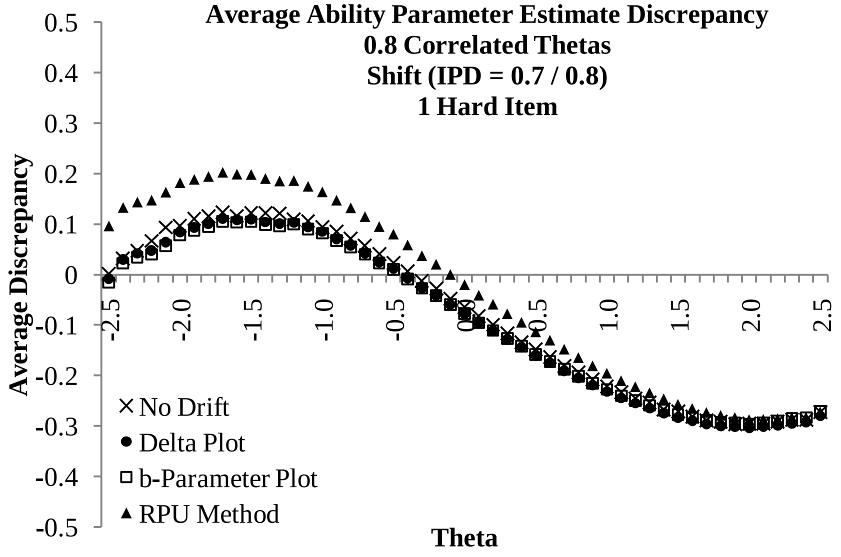
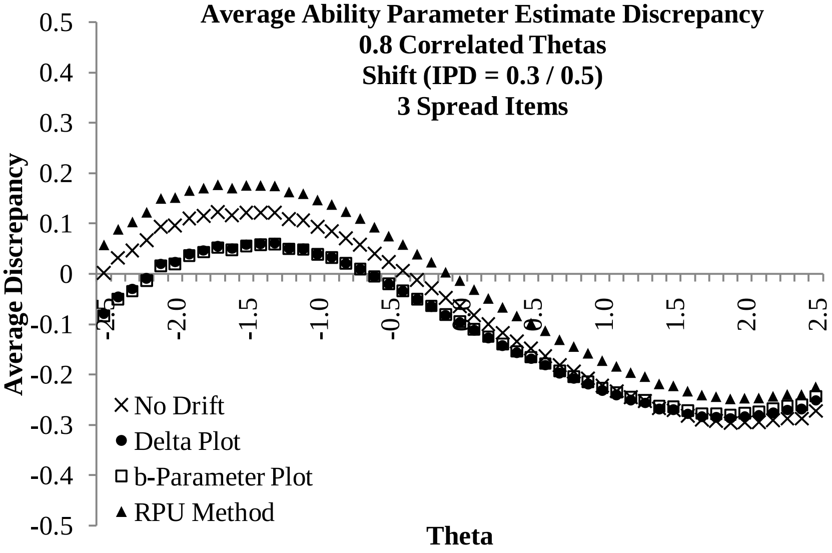


Figure C.1. Average Ability Parameter Estimate Discrepancy Plots

Continued, next page.

Figure C.1., cont'd.:



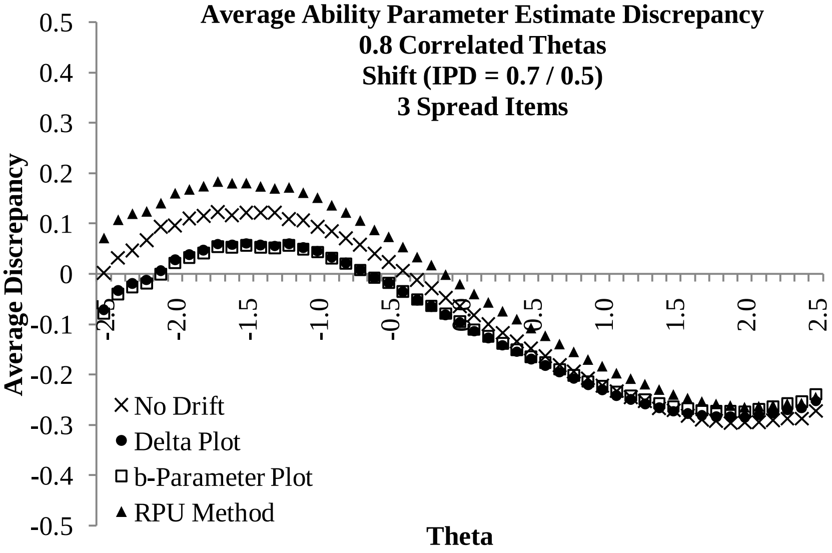
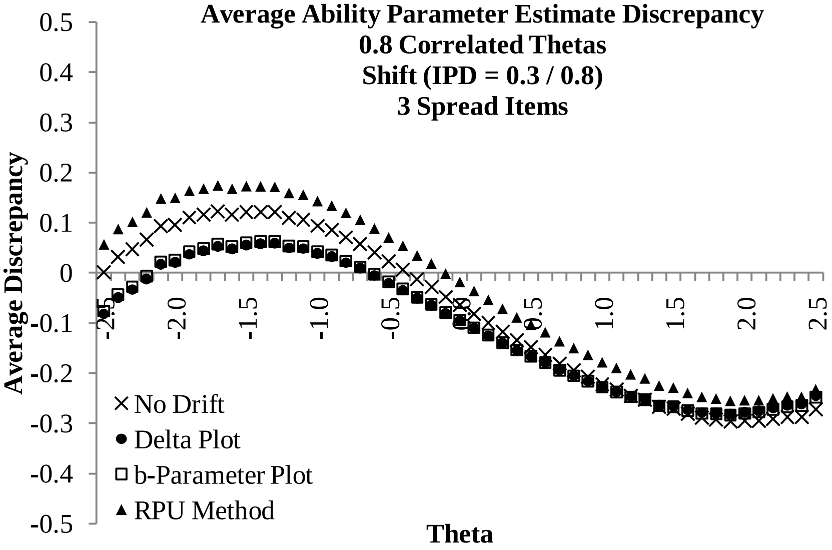


Figure C.1. Average Ability Parameter Estimate Discrepancy Plots

Continued, next page.

Figure C.1., cont'd.:



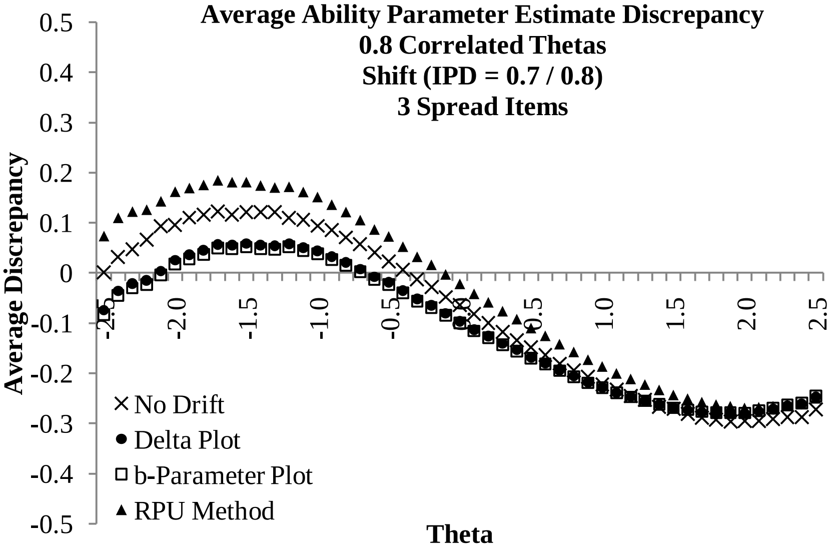
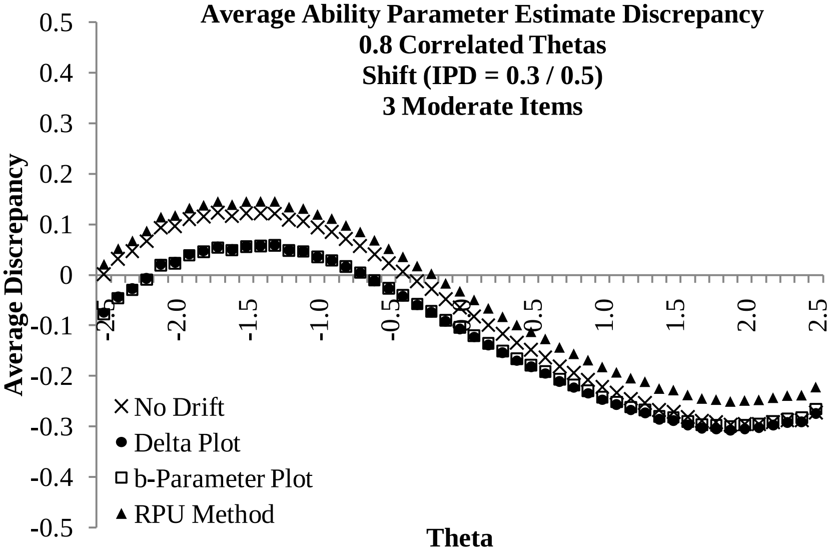


Figure C.1. Average Ability Parameter Estimate Discrepancy Plots

Continued, next page.

Figure C.1., cont'd.:



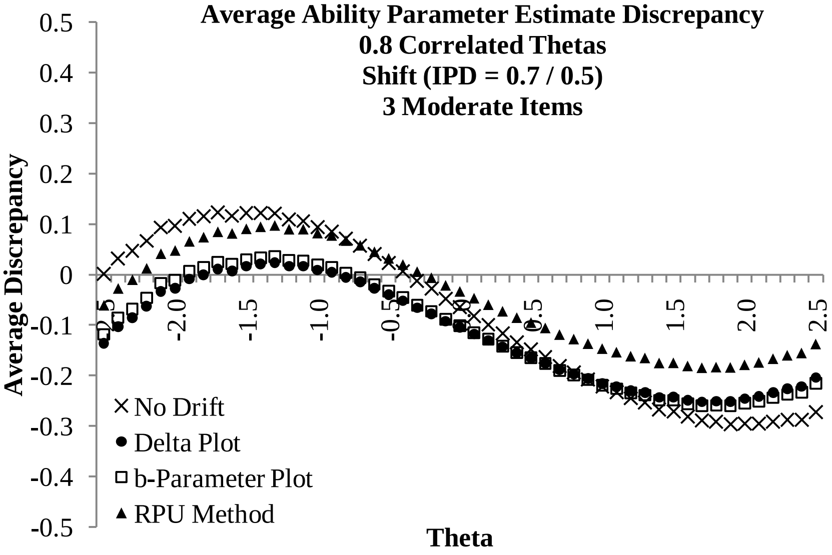
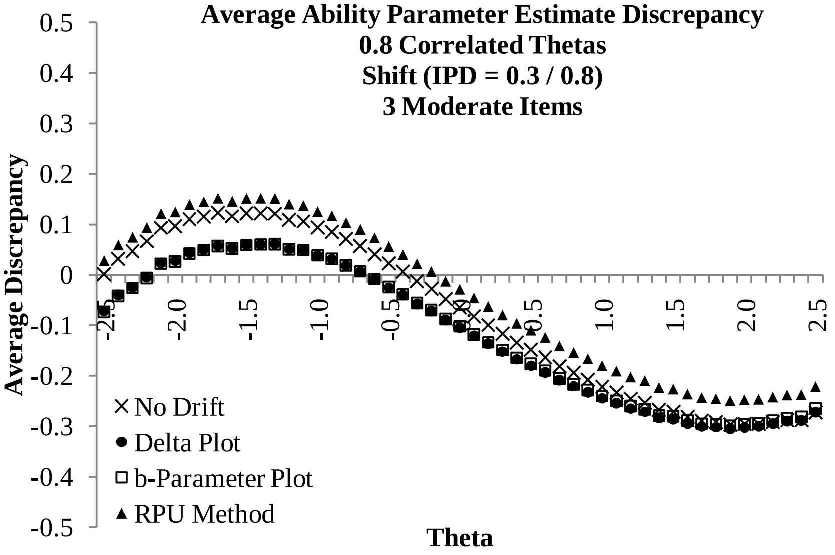


Figure C.1. Average Ability Parameter Estimate Discrepancy Plots

Continued, next page.

Figure C.1., cont'd.:



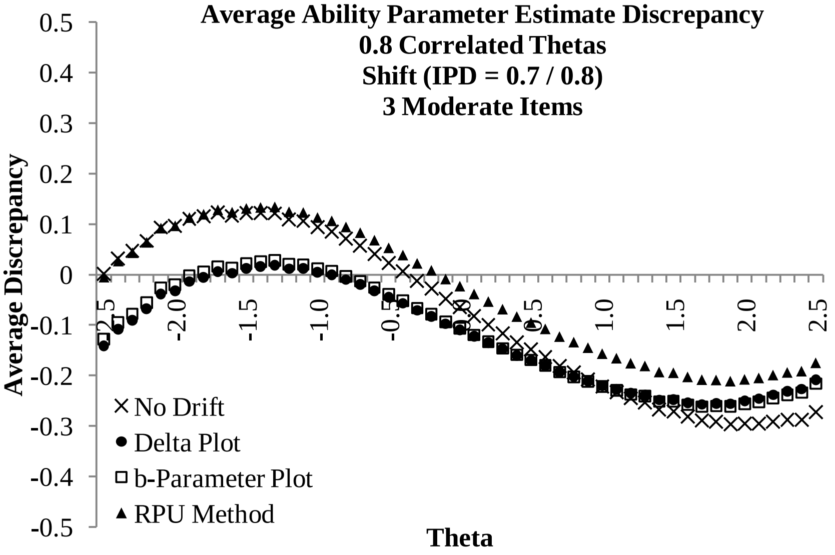
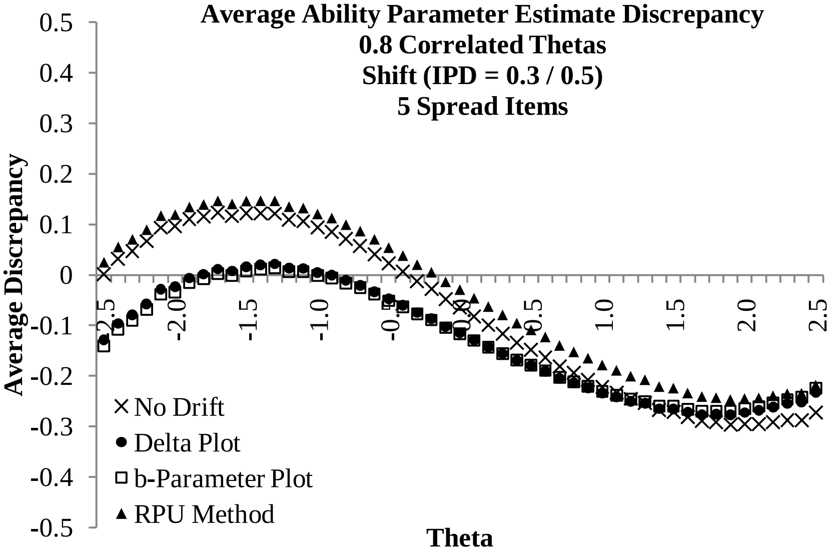


Figure C.1. Average Ability Parameter Estimate Discrepancy Plots

Continued, next page.

Figure C.1., cont'd.:



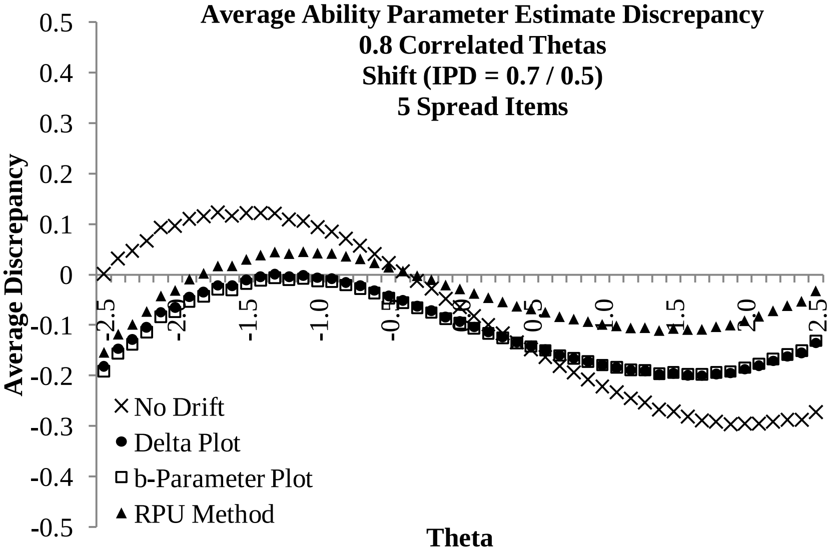


Figure C.1. Average Ability Parameter Estimate Discrepancy Plots

Continued, next page.

Figure C.1., cont'd.:

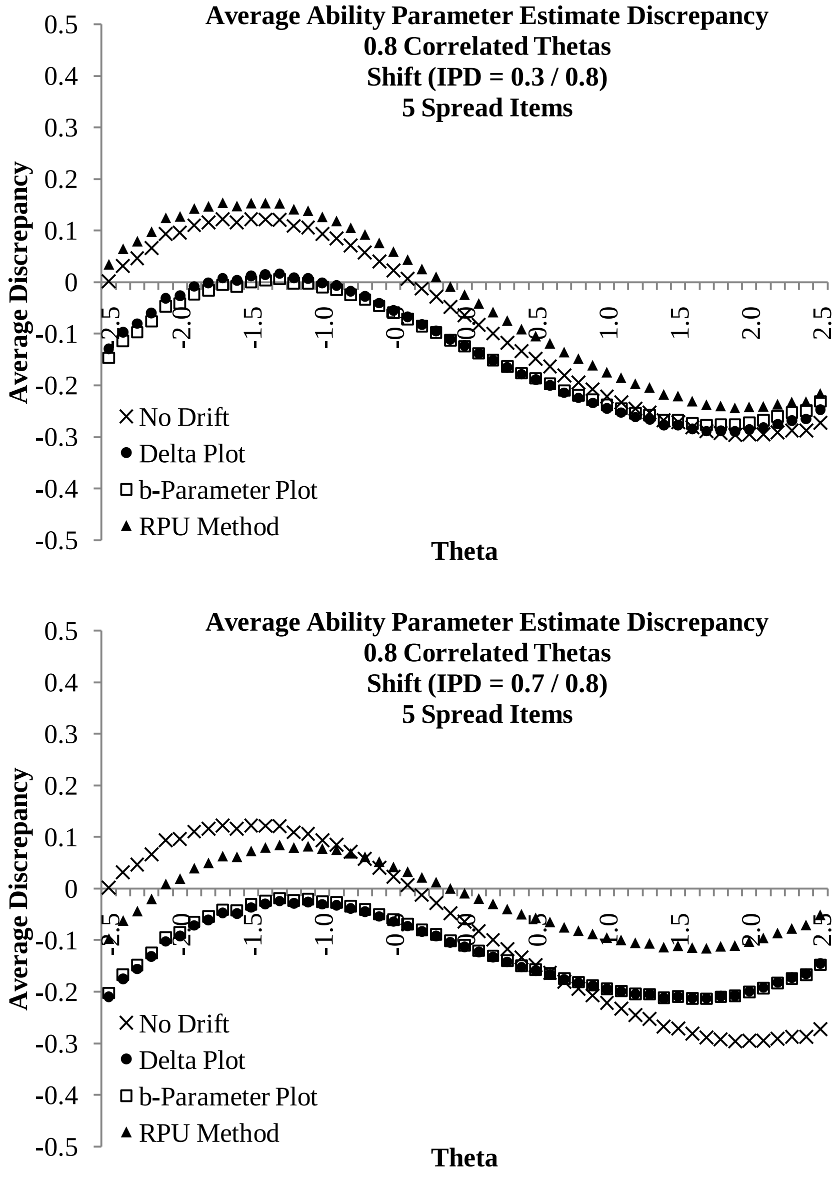
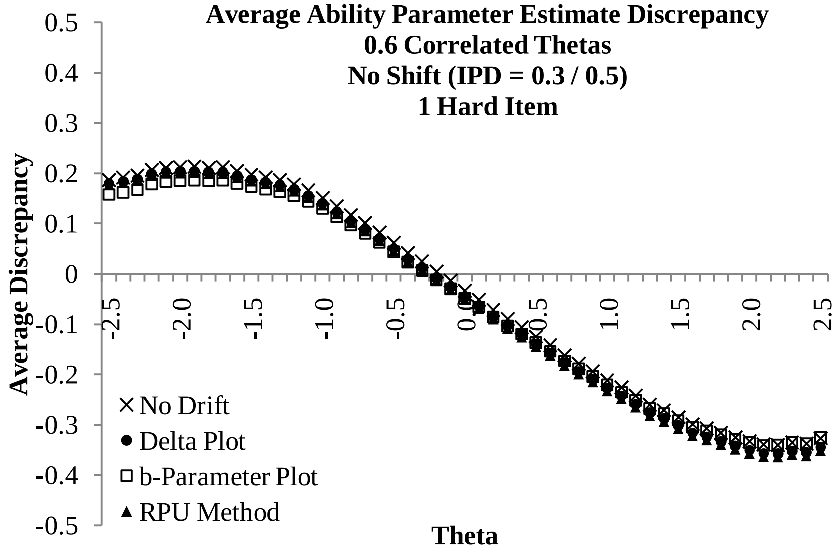


Figure C.1. Average Ability Parameter Estimate Discrepancy Plots

Continued, next page.

Figure C.1., cont'd.:



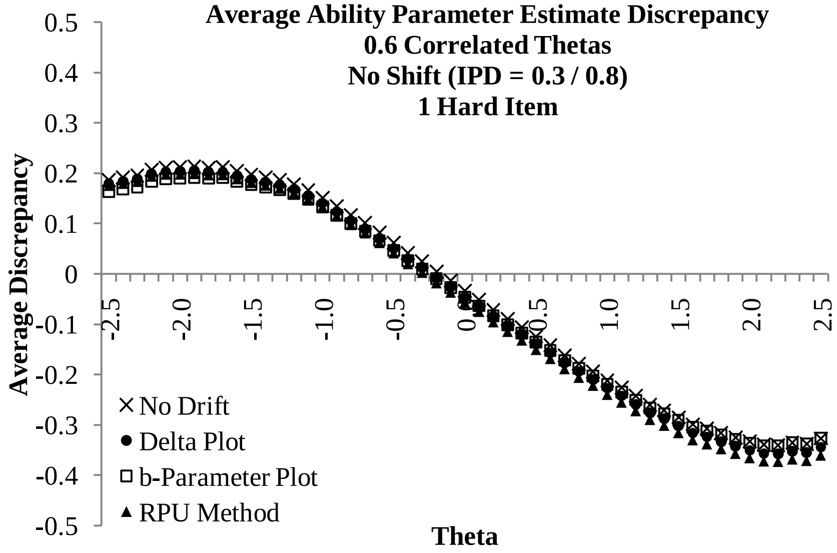
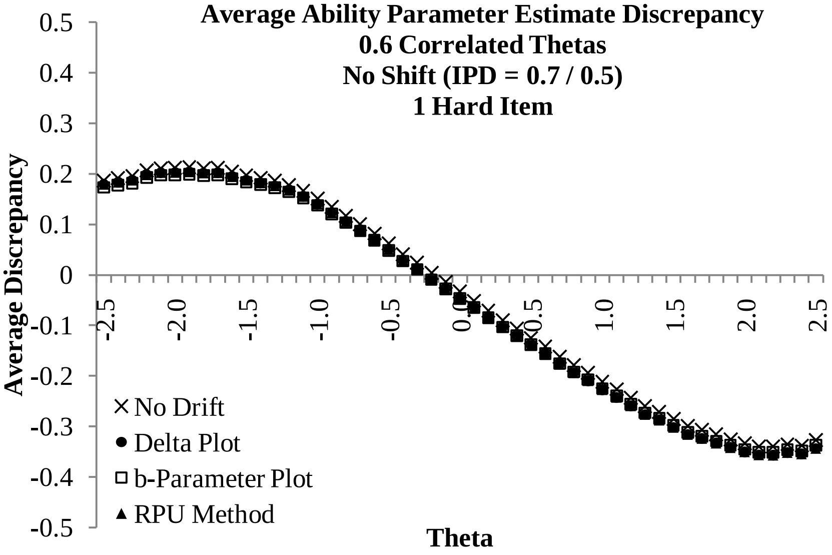


Figure C.1. Average Ability Parameter Estimate Discrepancy Plots

Continued, next page.

Figure C.1., cont'd.:



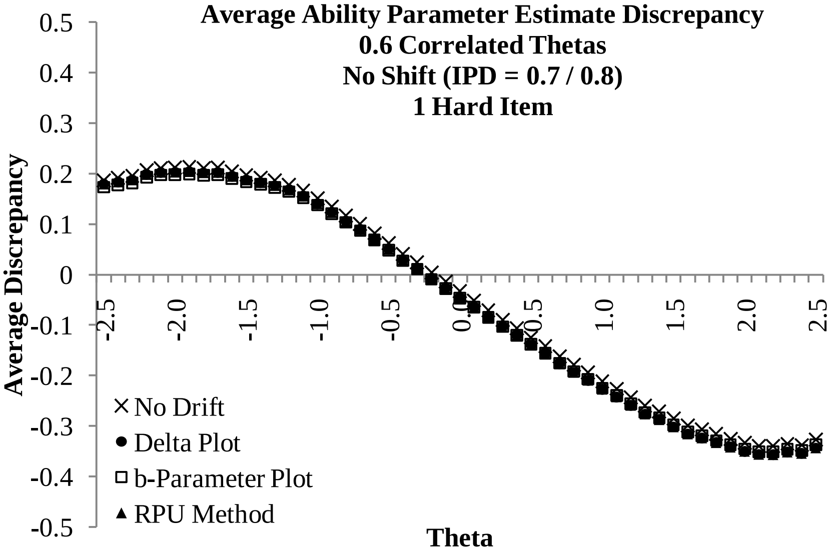
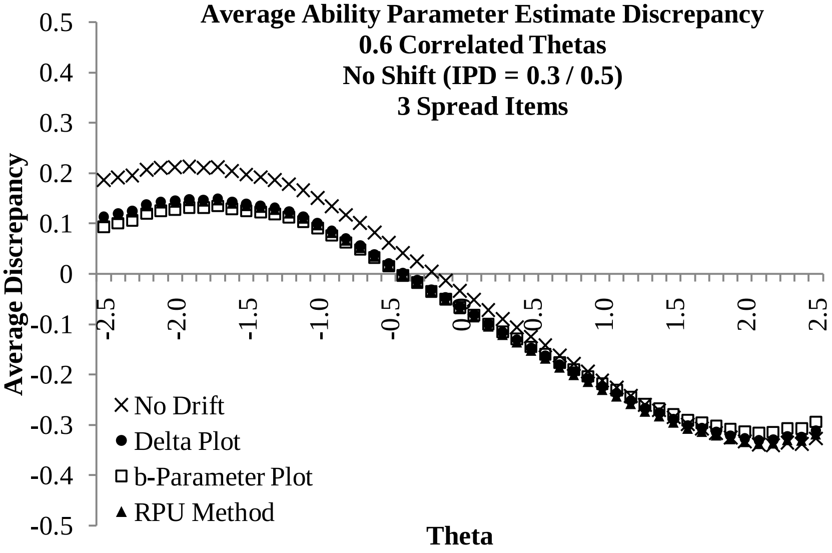


Figure C.1. Average Ability Parameter Estimate Discrepancy Plots

Continued, next page.

Figure C.1., cont'd.:



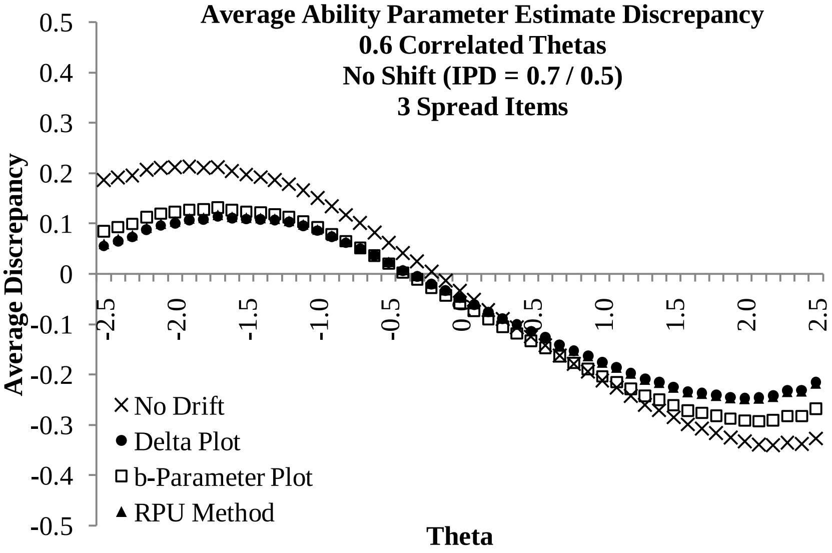
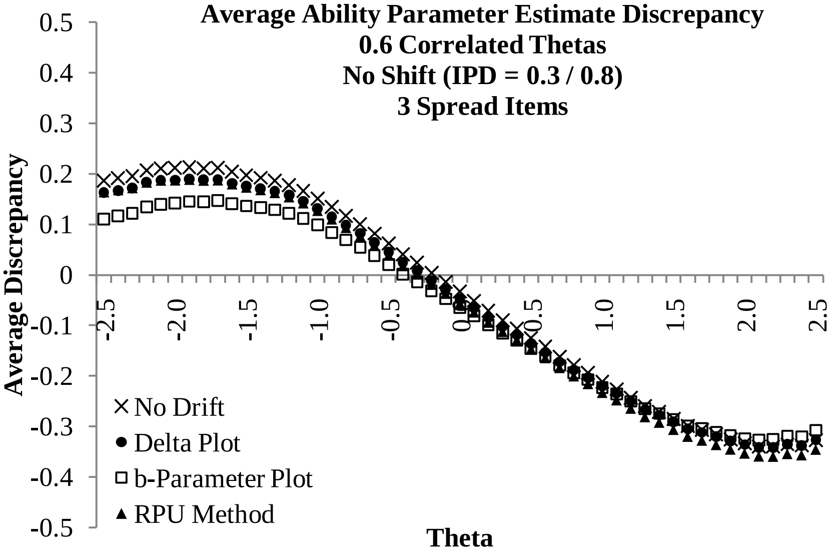


Figure C.1. Average Ability Parameter Estimate Discrepancy Plots

Continued, next page.

Figure C.1., cont'd.:



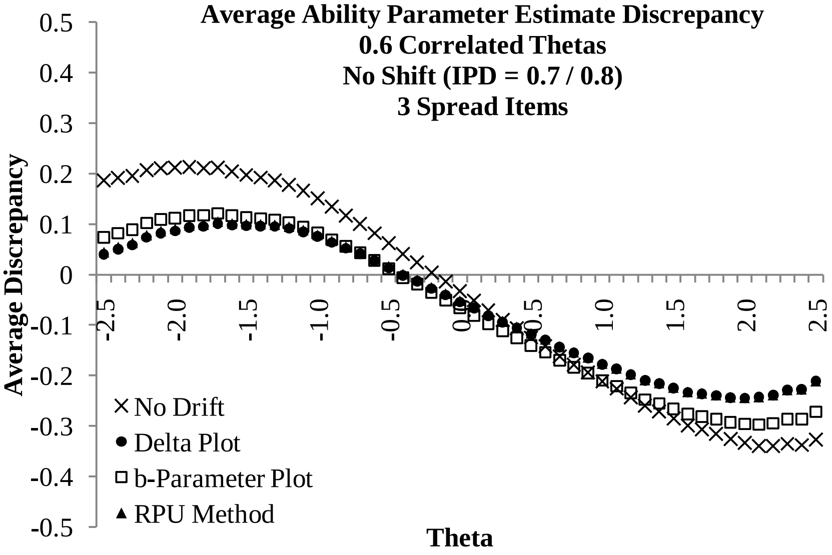
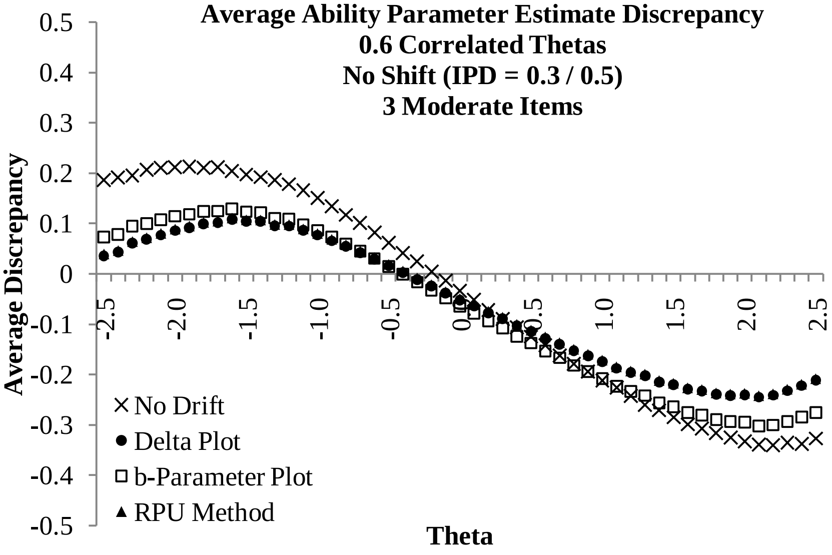


Figure C.1. Average Ability Parameter Estimate Discrepancy Plots

Continued, next page.

Figure C.1., cont'd.:



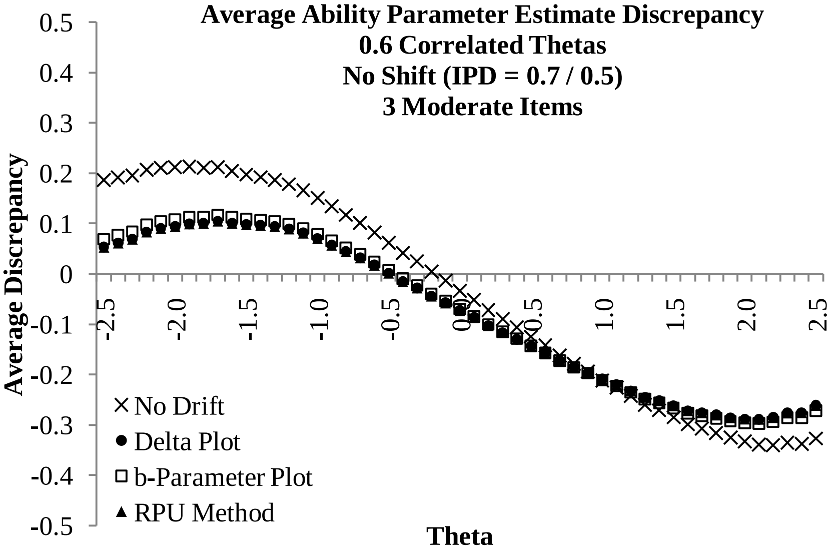
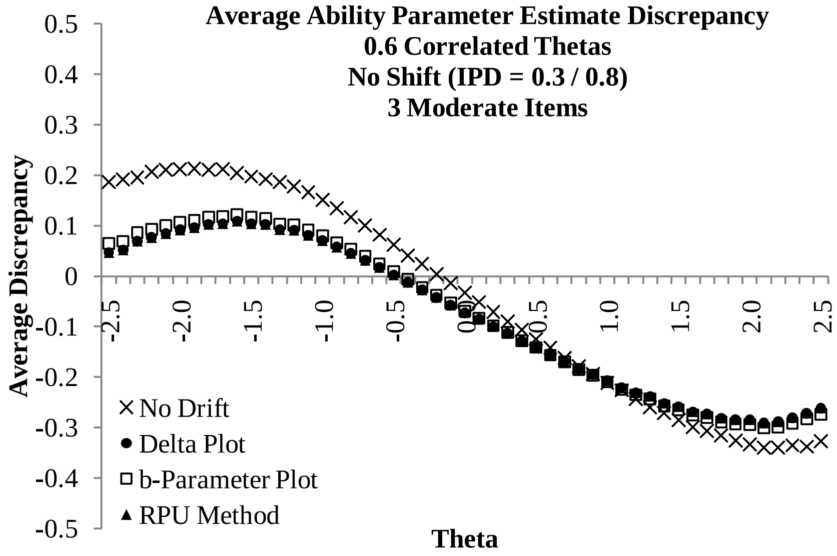


Figure C.1. Average Ability Parameter Estimate Discrepancy Plots

Continued, next page.

Figure C.1., cont'd.:



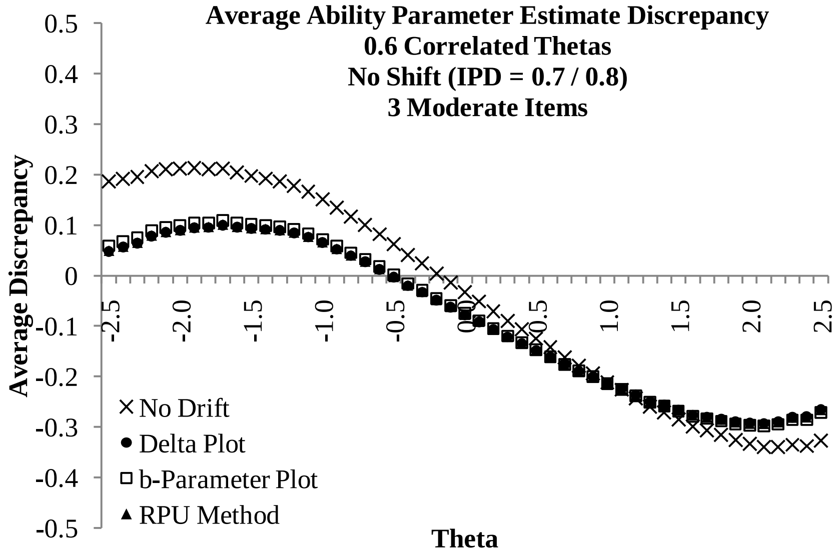
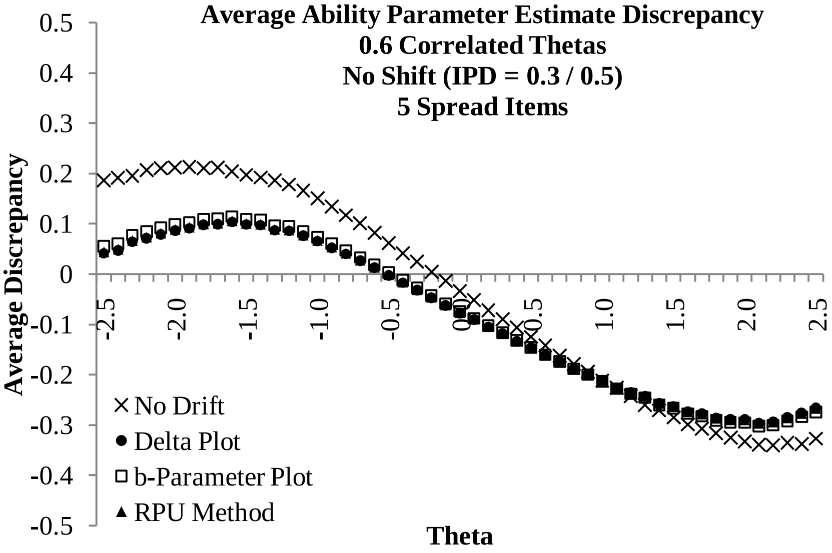


Figure C.1. Average Ability Parameter Estimate Discrepancy Plots

Continued, next page.

Figure C.1., cont'd.:



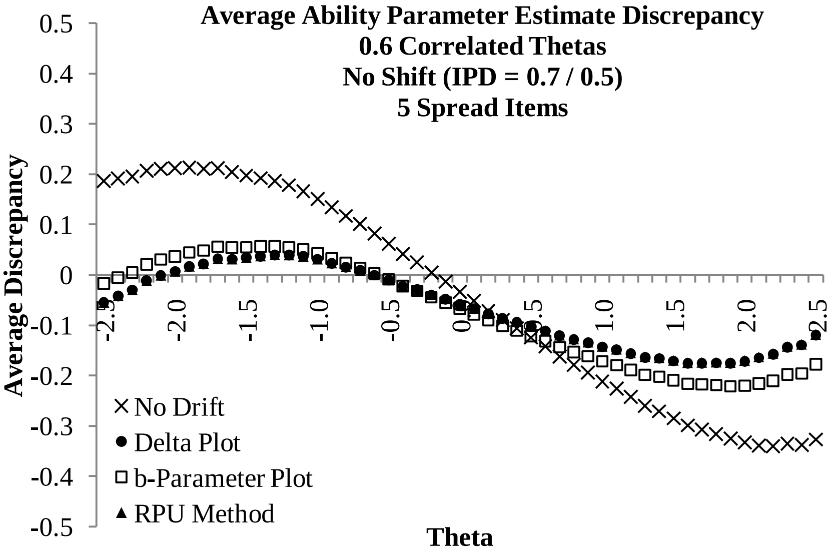
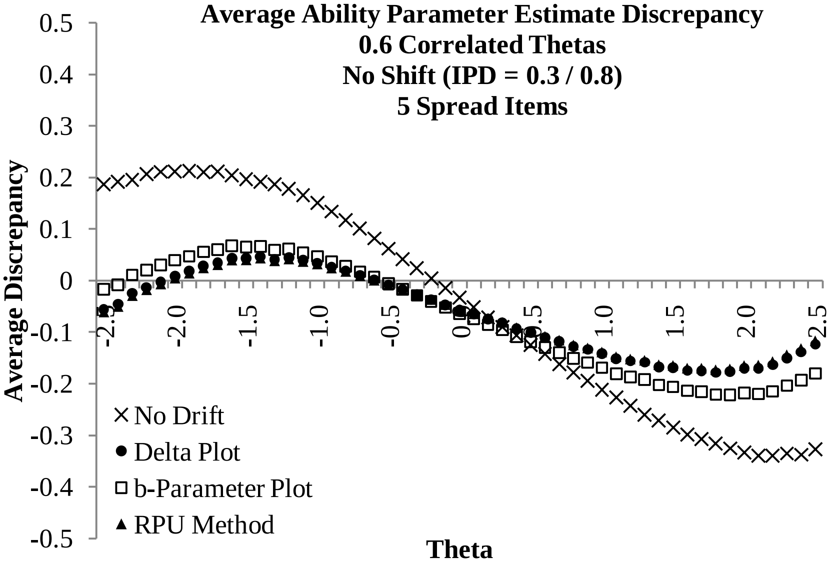


Figure C.1. Average Ability Parameter Estimate Discrepancy Plots

Continued, next page.

Figure C.1., cont'd.:



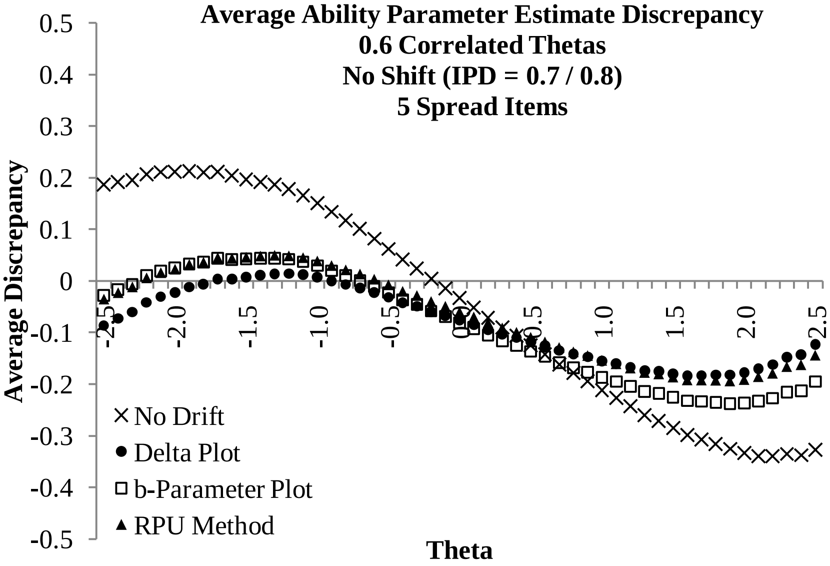
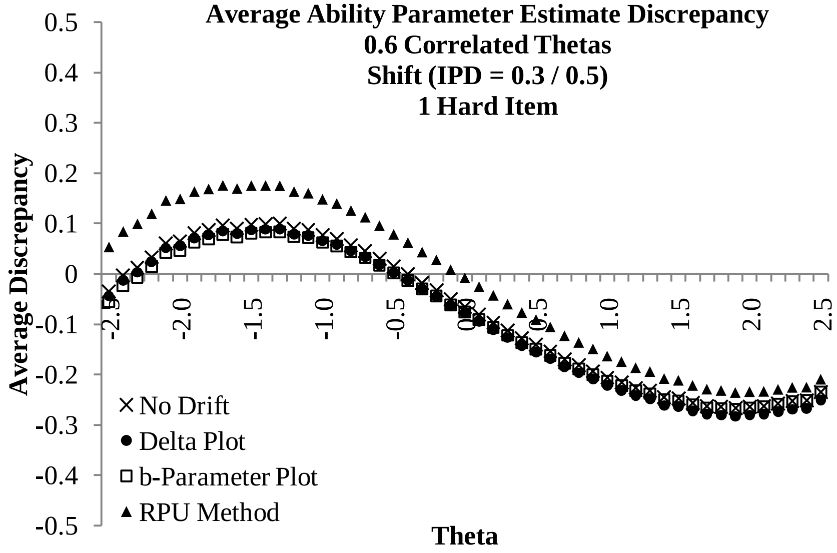


Figure C.1. Average Ability Parameter Estimate Discrepancy Plots

Continued, next page.

Figure C.1., cont'd.:



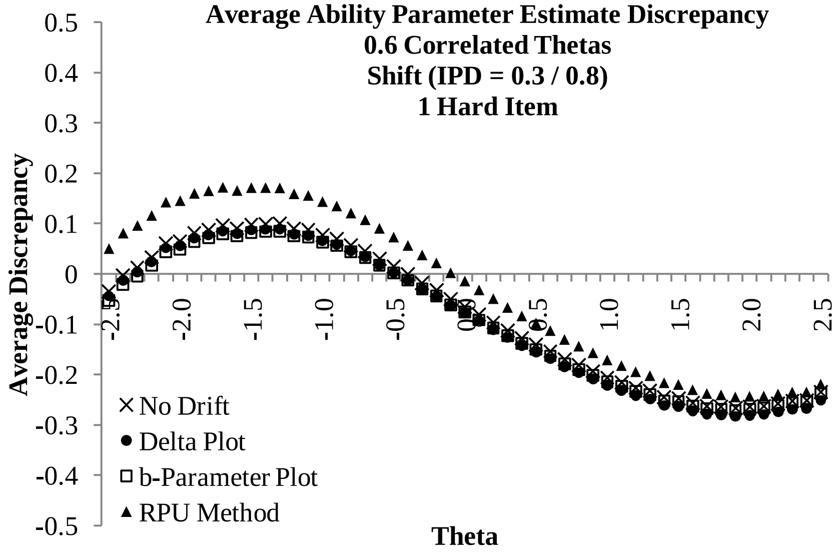
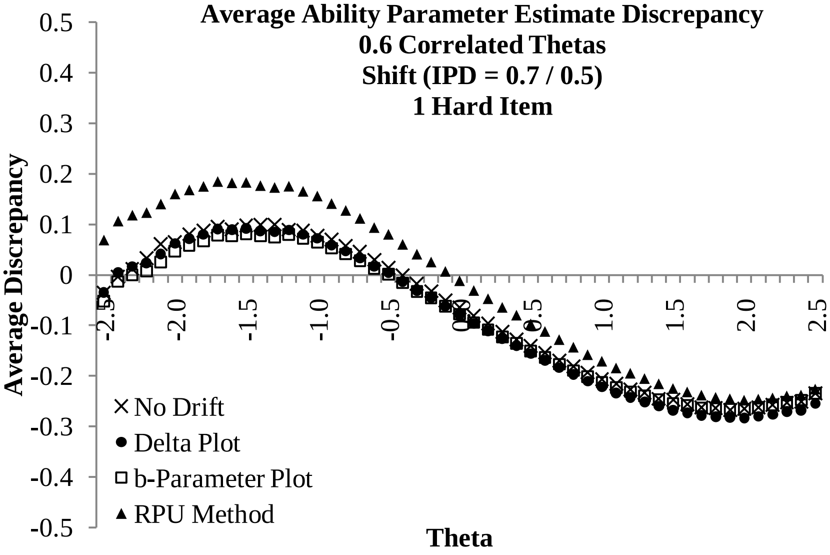


Figure C.1. Average Ability Parameter Estimate Discrepancy Plots

Continued, next page.

Figure C.1., cont'd.:



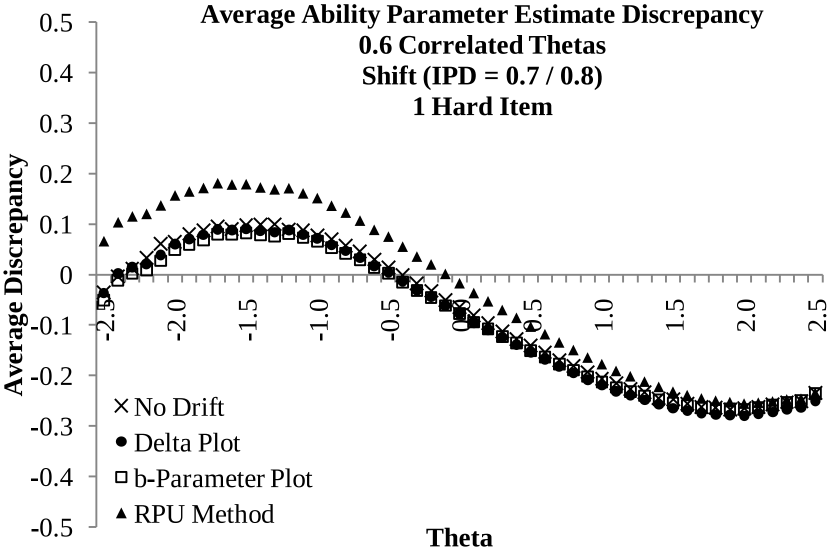
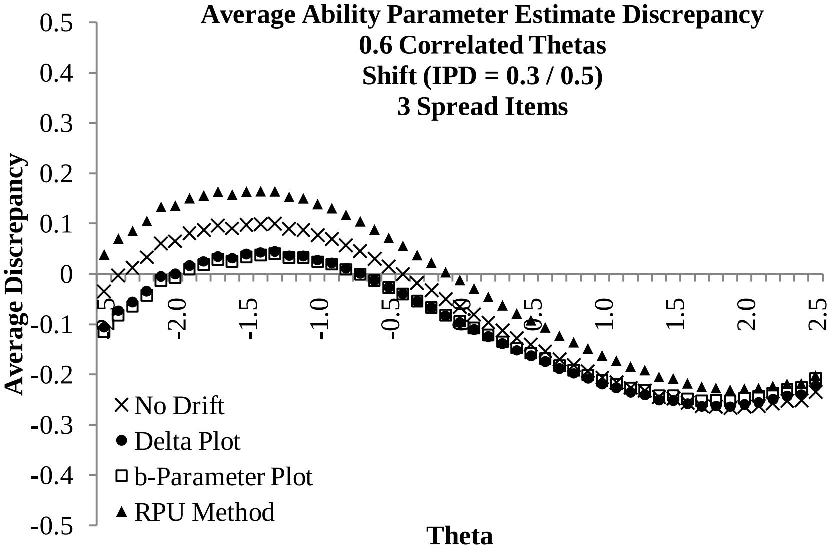


Figure C.1. Average Ability Parameter Estimate Discrepancy Plots

Continued, next page.

Figure C.1., cont'd.:



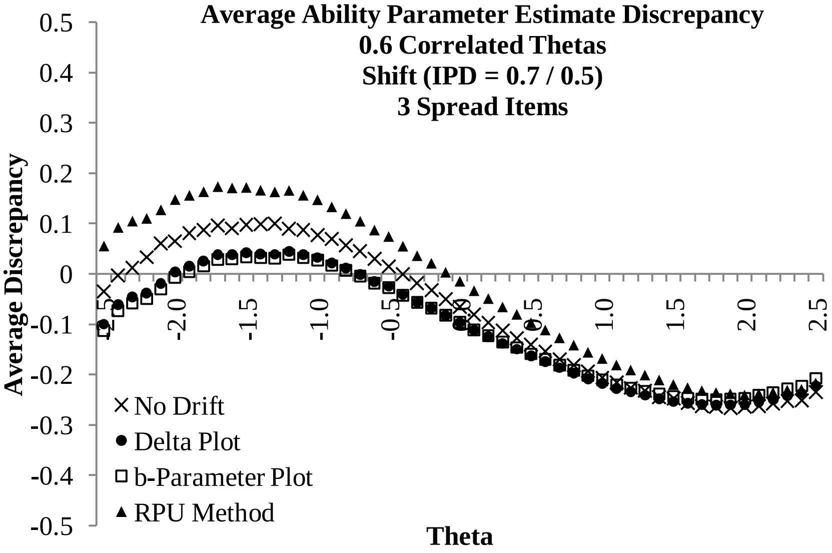
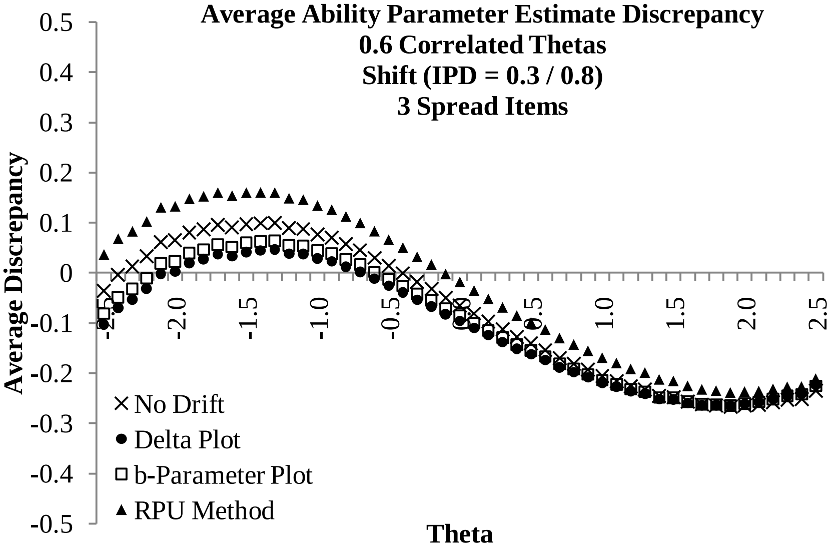


Figure C.1. Average Ability Parameter Estimate Discrepancy Plots

Continued, next page.

Figure C.1., cont'd.:



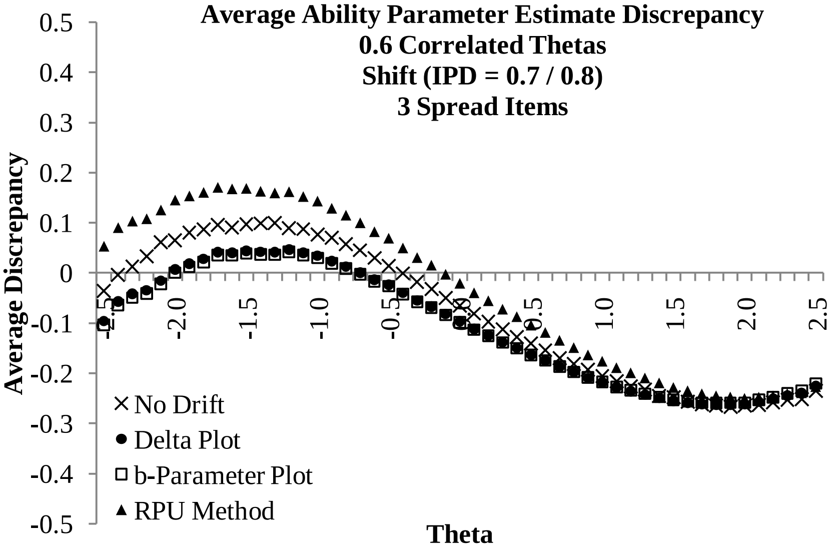
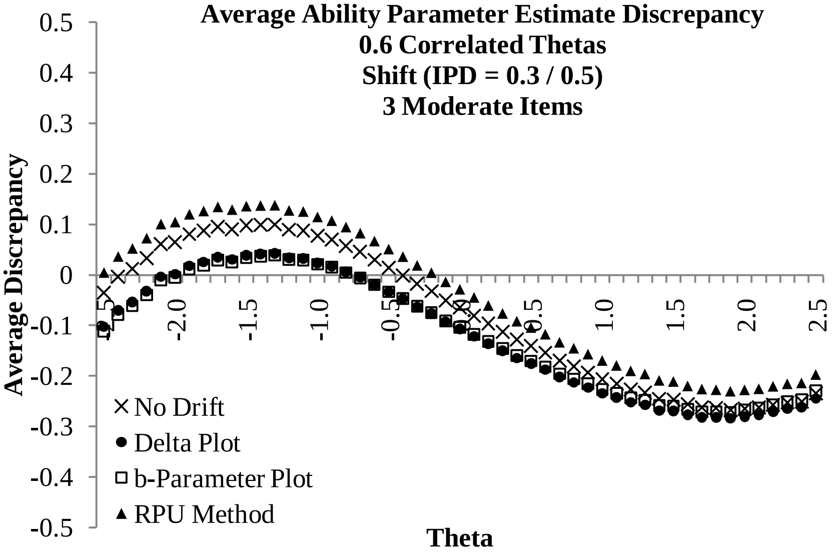


Figure C.1. Average Ability Parameter Estimate Discrepancy Plots

Continued, next page.

Figure C.1., cont'd.:



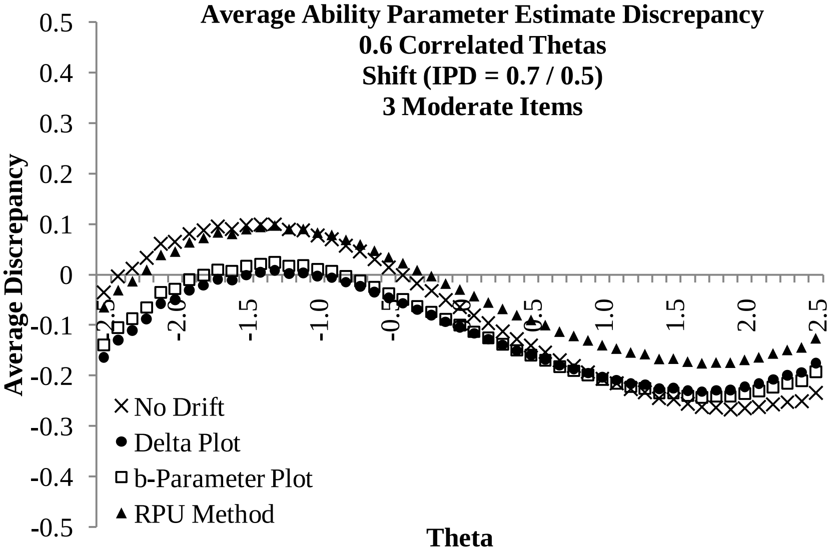
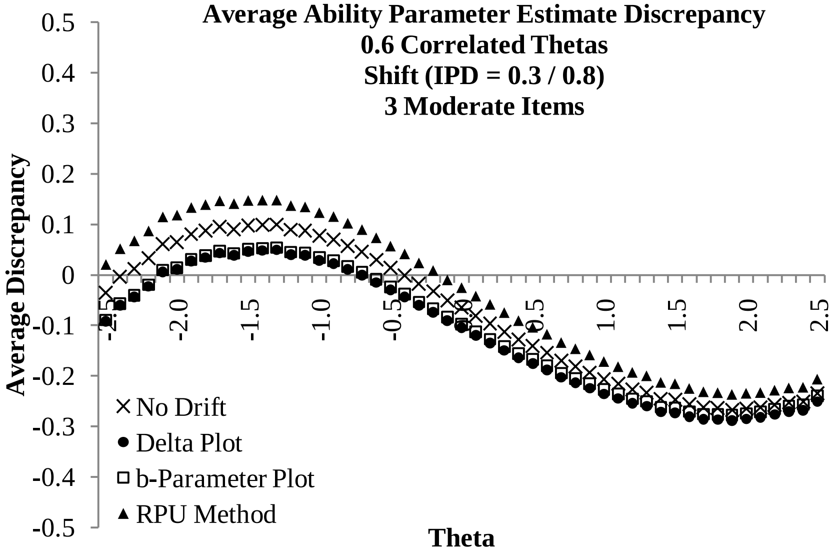


Figure C.1. Average Ability Parameter Estimate Discrepancy Plots

Continued, next page.

Figure C.1., cont'd.:



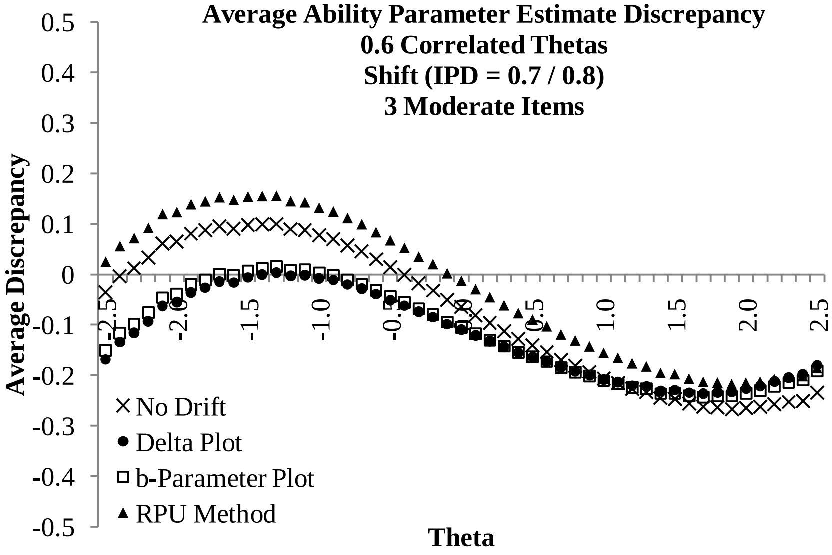
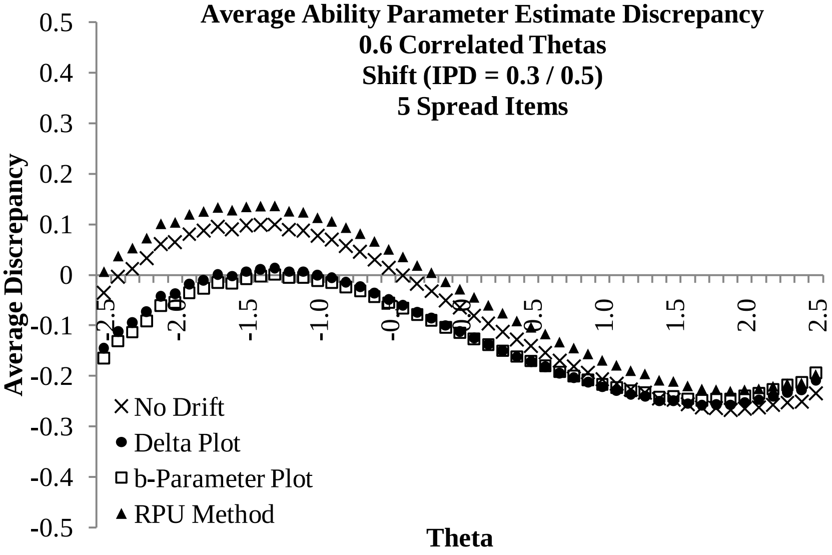


Figure C.1. Average Ability Parameter Estimate Discrepancy Plots

Continued, next page.

Figure C.1., cont'd.:



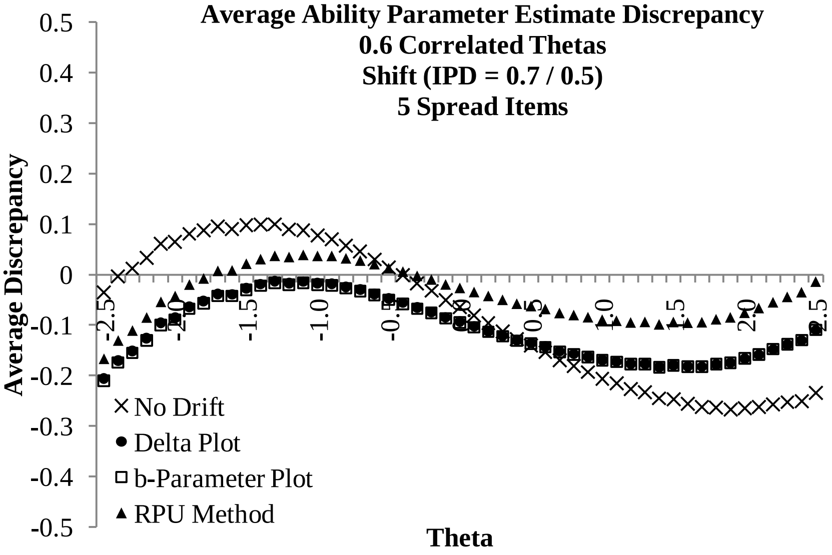
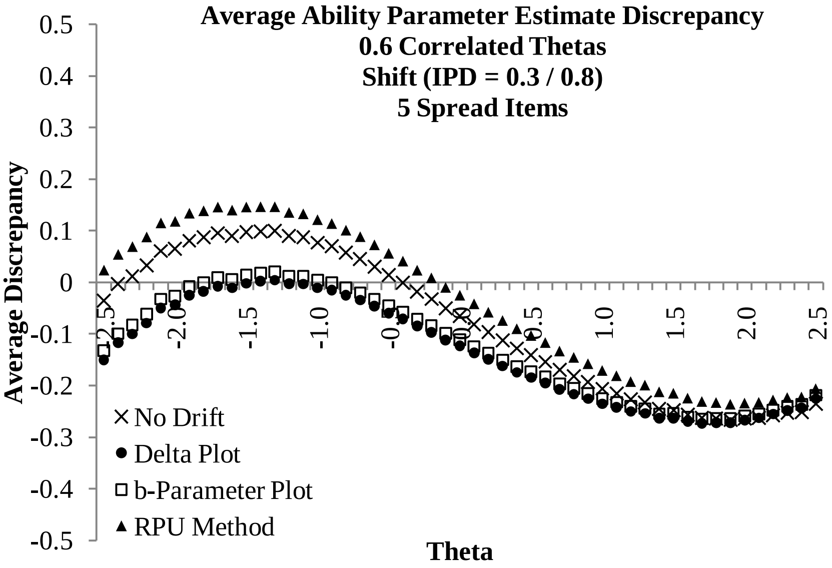


Figure C.1. Average Ability Parameter Estimate Discrepancy Plots

Continued, next page.

Figure C.1., cont'd.:



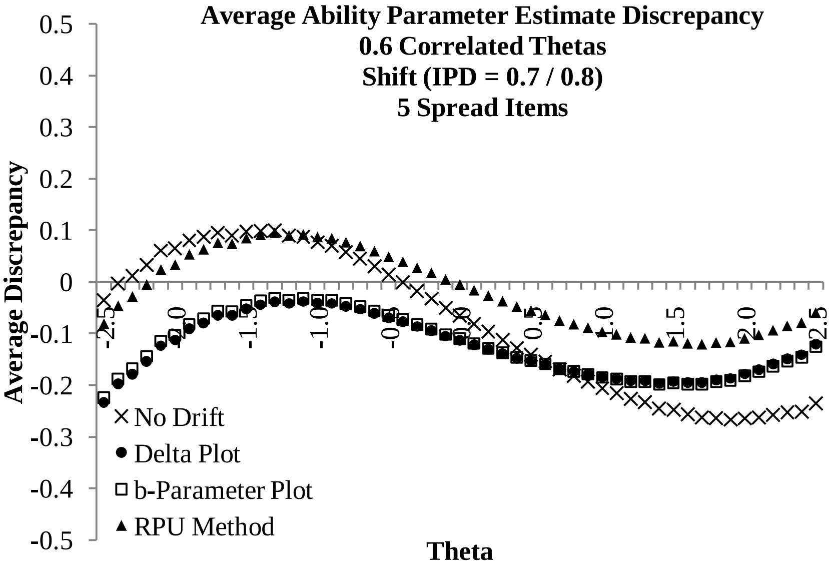


Figure C.1. Average Ability Parameter Estimate Discrepancy Plots

APPENDIX D

PERFORMANCE LEVEL CONTINGENCY TABLES

Table D.1. Performance Level Contingency Tables

Performance Level Contingency Tables, True vs. Aberrant & Purified Classification, 1 Hard Item with -0.3 *a* & -0.5 *b* Parameter Drift, Administration 1 Correlated Thetas of 0.8, Administration 2 Normal (No Shift) Distribution

		Aberrant Classification			
		1	2	3	4
True Classification	1	0.205	0.022	0.000	0.000
	2	0.050	0.171	0.053	0.000
	3	0.000	0.030	0.342	0.056
	4	0.000	0.000	0.003	0.067

		Delta Plot Purified Classification			
		1	2	3	4
True Classification	1	0.205	0.022	0.000	0.000
	2	0.051	0.171	0.052	0.000
	3	0.000	0.031	0.343	0.055
	4	0.000	0.000	0.003	0.067

		b Plot Purified Classification			
		1	2	3	4
True Classification	1	0.205	0.022	0.000	0.000
	2	0.050	0.172	0.052	0.000
	3	0.000	0.031	0.344	0.053
	4	0.000	0.000	0.003	0.067

		RPU Purified Classification			
		1	2	3	4
True Classification	1	0.205	0.022	0.000	0.000
	2	0.050	0.171	0.053	0.000
	3	0.000	0.030	0.342	0.056
	4	0.000	0.000	0.003	0.067

Continued, next page.

Table D.1., cont'd.:

Performance Level Contingency Tables, True vs. Aberrant & Purified Classification, 1 Hard Item with -0.7 *a* & -0.5 *b* Parameter Drift, Administration 1 Correlated Thetas of 0.8, Administration 2 Normal (No Shift) Distribution

		Aberrant Classification			
		1	2	3	4
True Classification	1	0.199	0.029	0.000	0.000
	2	0.040	0.170	0.064	0.000
	3	0.000	0.024	0.344	0.061
	4	0.000	0.000	0.002	0.068

		Delta Plot Purified Classification			
		1	2	3	4
True Classification	1	0.205	0.022	0.000	0.000
	2	0.051	0.171	0.052	0.000
	3	0.000	0.031	0.343	0.055
	4	0.000	0.000	0.003	0.067

		b Plot Purified Classification			
		1	2	3	4
True Classification	1	0.205	0.022	0.000	0.000
	2	0.051	0.171	0.052	0.000
	3	0.000	0.031	0.343	0.054
	4	0.000	0.000	0.003	0.067

		RPU Purified Classification			
		1	2	3	4
True Classification	1	0.205	0.022	0.000	0.000
	2	0.051	0.171	0.052	0.000
	3	0.000	0.031	0.343	0.055
	4	0.000	0.000	0.003	0.067

Continued, next page.

Table D.1., cont'd.:

Performance Level Contingency Tables, True vs. Aberrant & Purified Classification, 1 Hard Item with -0.3 *a* & -0.8 *b* Parameter Drift, Administration 1 Correlated Thetas of 0.8, Administration 2 Normal (No Shift) Distribution

		Aberrant Classification			
		1	2	3	4
True Classification	1	0.205	0.023	0.000	0.000
	2	0.050	0.170	0.054	0.000
	3	0.000	0.029	0.342	0.057
	4	0.000	0.000	0.003	0.067

		Delta Plot Purified Classification			
		1	2	3	4
True Classification	1	0.205	0.022	0.000	0.000
	2	0.051	0.171	0.052	0.000
	3	0.000	0.031	0.343	0.055
	4	0.000	0.000	0.003	0.067

		b Plot Purified Classification			
		1	2	3	4
True Classification	1	0.205	0.022	0.000	0.000
	2	0.050	0.172	0.052	0.000
	3	0.000	0.031	0.344	0.054
	4	0.000	0.000	0.003	0.067

		RPU Purified Classification			
		1	2	3	4
True Classification	1	0.205	0.023	0.000	0.000
	2	0.050	0.170	0.054	0.000
	3	0.000	0.029	0.342	0.057
	4	0.000	0.000	0.003	0.067

Continued, next page.

Table D.1., cont’d.:

Performance Level Contingency Tables, True vs. Aberrant & Purified Classification, 1 Hard Item with -0.7 *a* & -0.8 *b* Parameter Drift, Administration 1 Correlated Thetas of 0.8, Administration 2 Normal (No Shift) Distribution

		Aberrant Classification			
		1	2	3	4
True Classification	1	0.199	0.029	0.000	0.000
	2	0.040	0.170	0.064	0.000
	3	0.000	0.024	0.344	0.061
	4	0.000	0.000	0.002	0.068

		Delta Plot Purified Classification			
		1	2	3	4
True Classification	1	0.205	0.022	0.000	0.000
	2	0.051	0.171	0.052	0.000
	3	0.000	0.031	0.343	0.055
	4	0.000	0.000	0.003	0.067

		b Plot Purified Classification			
		1	2	3	4
True Classification	1	0.205	0.022	0.000	0.000
	2	0.051	0.171	0.052	0.000
	3	0.000	0.031	0.343	0.054
	4	0.000	0.000	0.003	0.067

		RPU Purified Classification			
		1	2	3	4
True Classification	1	0.205	0.022	0.000	0.000
	2	0.051	0.171	0.052	0.000
	3	0.000	0.031	0.343	0.055
	4	0.000	0.000	0.003	0.067

Continued, next page.

Table D.1., cont'd.:

Performance Level Contingency Tables, True vs. Aberrant & Purified Classification, 3 Spread Items with -0.3 *a* & -0.5 *b* Parameter Drift, Administration 1 Correlated Thetas of 0.8, Administration 2 Normal (No Shift) Distribution

		Aberrant Classification			
		1	2	3	4
True Classification	1	0.202	0.025	0.000	0.000
	2	0.045	0.172	0.057	0.000
	3	0.000	0.028	0.346	0.055
	4	0.000	0.000	0.003	0.067

		Delta Plot Purified Classification			
		1	2	3	4
True Classification	1	0.202	0.025	0.000	0.000
	2	0.046	0.173	0.056	0.000
	3	0.000	0.028	0.347	0.053
	4	0.000	0.000	0.003	0.067

		b Plot Purified Classification			
		1	2	3	4
True Classification	1	0.202	0.026	0.000	0.000
	2	0.045	0.173	0.056	0.000
	3	0.000	0.028	0.348	0.052
	4	0.000	0.000	0.003	0.067

		RPU Purified Classification			
		1	2	3	4
True Classification	1	0.202	0.025	0.000	0.000
	2	0.045	0.172	0.057	0.000
	3	0.000	0.028	0.346	0.055
	4	0.000	0.000	0.003	0.067

Continued, next page.

Table D.1., cont’d.:

Performance Level Contingency Tables, True vs. Aberrant & Purified Classification, 3 Spread Items with -0.7 *a* & -0.5 *b* Parameter Drift, Administration 1 Correlated Thetas of 0.8, Administration 2 Normal (No Shift) Distribution

		Aberrant Classification			
		1	2	3	4
True Classification	1	0.196	0.032	0.000	0.000
	2	0.037	0.175	0.062	0.000
	3	0.000	0.025	0.354	0.050
	4	0.000	0.000	0.003	0.066

		Delta Plot Purified Classification			
		1	2	3	4
True Classification	1	0.202	0.026	0.000	0.000
	2	0.045	0.176	0.053	0.000
	3	0.000	0.031	0.352	0.045
	4	0.000	0.000	0.004	0.066

		b Plot Purified Classification			
		1	2	3	4
True Classification	1	0.202	0.026	0.000	0.000
	2	0.045	0.174	0.055	0.000
	3	0.000	0.029	0.349	0.050
	4	0.000	0.000	0.003	0.066

		RPU Purified Classification			
		1	2	3	4
True Classification	1	0.202	0.025	0.000	0.000
	2	0.045	0.176	0.052	0.000
	3	0.000	0.031	0.352	0.045
	4	0.000	0.000	0.004	0.066

Continued, next page.

Table D.1., cont'd.:

Performance Level Contingency Tables, True vs. Aberrant & Purified Classification, 3 Spread Items with -0.7 *a* & -0.5 *b* Parameter Drift, Administration 1 Correlated Thetas of 0.8, Administration 2 Normal (No Shift) Distribution

		Aberrant Classification			
		1	2	3	4
True Classification	1	0.200	0.027	0.000	0.000
	2	0.042	0.172	0.060	0.000
	3	0.000	0.026	0.346	0.057
	4	0.000	0.000	0.003	0.067

		Delta Plot Purified Classification			
		1	2	3	4
True Classification	1	0.204	0.023	0.000	0.000
	2	0.049	0.172	0.053	0.000
	3	0.000	0.030	0.346	0.052
	4	0.000	0.000	0.003	0.067

		b Plot Purified Classification			
		1	2	3	4
True Classification	1	0.201	0.026	0.000	0.000
	2	0.044	0.173	0.057	0.000
	3	0.000	0.028	0.347	0.053
	4	0.000	0.000	0.003	0.067

		RPU Purified Classification			
		1	2	3	4
True Classification	1	0.204	0.023	0.000	0.000
	2	0.049	0.171	0.054	0.000
	3	0.000	0.030	0.344	0.055
	4	0.000	0.000	0.003	0.067

Continued, next page.

Table D.1., cont'd.:

Performance Level Contingency Tables, True vs. Aberrant & Purified Classification, 3 Spread Items with -0.7 *a* & -0.8 *b* Parameter Drift, Administration 1 Correlated Thetas of 0.8, Administration 2 Normal (No Shift) Distribution

		Aberrant Classification			
		1	2	3	4
True Classification	1	0.195	0.033	0.000	0.000
	2	0.035	0.175	0.064	0.000
	3	0.000	0.024	0.355	0.050
	4	0.000	0.000	0.003	0.066

		Delta Plot Purified Classification			
		1	2	3	4
True Classification	1	0.201	0.027	0.000	0.000
	2	0.044	0.177	0.054	0.000
	3	0.000	0.030	0.353	0.045
	4	0.000	0.000	0.004	0.066

		b Plot Purified Classification			
		1	2	3	4
True Classification	1	0.201	0.027	0.000	0.000
	2	0.043	0.174	0.057	0.000
	3	0.000	0.028	0.350	0.051
	4	0.000	0.000	0.003	0.066

		RPU Purified Classification			
		1	2	3	4
True Classification	1	0.201	0.026	0.000	0.000
	2	0.044	0.177	0.053	0.000
	3	0.000	0.030	0.353	0.045
	4	0.000	0.000	0.004	0.066

Continued, next page.

Table D.1., cont'd.:

Performance Level Contingency Tables, True vs. Aberrant & Purified Classification, 3 Moderate Items with -0.3 *a* & -0.5 *b* Parameter Drift, Administration 1 Correlated Thetas of 0.8, Administration 2 Normal (No Shift) Distribution

		Aberrant Classification			
		1	2	3	4
True Classification	1	0.195	0.032	0.000	0.000
	2	0.036	0.175	0.063	0.000
	3	0.000	0.024	0.355	0.049
	4	0.000	0.000	0.004	0.066

		Delta Plot Purified Classification			
		1	2	3	4
True Classification	1	0.201	0.026	0.000	0.000
	2	0.044	0.177	0.053	0.000
	3	0.000	0.030	0.353	0.045
	4	0.000	0.000	0.004	0.066

		b Plot Purified Classification			
		1	2	3	4
True Classification	1	0.201	0.026	0.000	0.000
	2	0.044	0.174	0.056	0.000
	3	0.000	0.029	0.349	0.051
	4	0.000	0.000	0.003	0.066

		RPU Purified Classification			
		1	2	3	4
True Classification	1	0.201	0.026	0.000	0.000
	2	0.045	0.177	0.052	0.000
	3	0.000	0.031	0.353	0.045
	4	0.000	0.000	0.004	0.065

Continued, next page.

Table D.1., cont’d.:

Performance Level Contingency Tables, True vs. Aberrant & Purified Classification, 3 Moderate Items with -0.7 *a* & -0.5 *b* Parameter Drift, Administration 1 Correlated Thetas of 0.8, Administration 2 Normal (No Shift) Distribution

		Aberrant Classification			
		1	2	3	4
True Classification	1	0.194	0.034	0.000	0.000
	2	0.034	0.172	0.068	0.000
	3	0.000	0.022	0.351	0.055
	4	0.000	0.000	0.003	0.067

		Delta Plot Purified Classification			
		1	2	3	4
True Classification	1	0.200	0.028	0.000	0.000
	2	0.042	0.175	0.057	0.000
	3	0.000	0.027	0.351	0.050
	4	0.000	0.000	0.003	0.066

		b Plot Purified Classification			
		1	2	3	4
True Classification	1	0.200	0.027	0.000	0.000
	2	0.042	0.174	0.057	0.000
	3	0.000	0.028	0.350	0.051
	4	0.000	0.000	0.003	0.066

		RPU Purified Classification			
		1	2	3	4
True Classification	1	0.200	0.028	0.000	0.000
	2	0.042	0.175	0.057	0.000
	3	0.000	0.027	0.351	0.050
	4	0.000	0.000	0.003	0.066

Continued, next page.

Table D.1., cont'd.:

Performance Level Contingency Tables, True vs. Aberrant & Purified Classification, 3 Moderate Items with -0.7 *a* & -0.5 *b* Parameter Drift, Administration 1 Correlated Thetas of 0.8, Administration 2 Normal (No Shift) Distribution

		Aberrant Classification			
		1	2	3	4
True Classification	1	0.194	0.034	0.000	0.000
	2	0.034	0.173	0.067	0.000
	3	0.000	0.022	0.351	0.055
	4	0.000	0.000	0.003	0.067

		Delta Plot Purified Classification			
		1	2	3	4
True Classification	1	0.200	0.027	0.000	0.000
	2	0.042	0.175	0.057	0.000
	3	0.000	0.028	0.351	0.050
	4	0.000	0.000	0.003	0.066

		b Plot Purified Classification			
		1	2	3	4
True Classification	1	0.200	0.027	0.000	0.000
	2	0.043	0.174	0.057	0.000
	3	0.000	0.028	0.350	0.051
	4	0.000	0.000	0.003	0.066

		RPU Purified Classification			
		1	2	3	4
True Classification	1	0.200	0.027	0.000	0.000
	2	0.042	0.175	0.057	0.000
	3	0.000	0.028	0.351	0.050
	4	0.000	0.000	0.003	0.066

Continued, next page.

Table D.1., cont'd.:

Performance Level Contingency Tables, True vs. Aberrant & Purified Classification, 3 Moderate Items with -0.7 *a* & -0.8 *b* Parameter Drift, Administration 1 Correlated Thetas of 0.8, Administration 2 Normal (No Shift) Distribution

		Aberrant Classification			
		1	2	3	4
True Classification	1	0.193	0.034	0.000	0.000
	2	0.033	0.172	0.069	0.000
	3	0.000	0.021	0.351	0.056
	4	0.000	0.000	0.003	0.067

		Delta Plot Purified Classification			
		1	2	3	4
True Classification	1	0.199	0.028	0.000	0.000
	2	0.041	0.174	0.058	0.000
	3	0.000	0.027	0.351	0.051
	4	0.000	0.000	0.003	0.066

		b Plot Purified Classification			
		1	2	3	4
True Classification	1	0.200	0.028	0.000	0.000
	2	0.041	0.174	0.058	0.000
	3	0.000	0.027	0.350	0.051
	4	0.000	0.000	0.003	0.066

		RPU Purified Classification			
		1	2	3	4
True Classification	1	0.199	0.028	0.000	0.000
	2	0.041	0.174	0.058	0.000
	3	0.000	0.027	0.351	0.051
	4	0.000	0.000	0.003	0.066

Continued, next page.

Table D.1., cont'd.:

Performance Level Contingency Tables, True vs. Aberrant & Purified Classification, 5 Spread Items with -0.3 *a* & -0.5 *b* Parameter Drift, Administration 1 Correlated Thetas of 0.8, Administration 2 Normal (No Shift) Distribution

		Aberrant Classification			
		1	2	3	4
True Classification	1	0.193	0.034	0.000	0.000
	2	0.034	0.172	0.068	0.000
	3	0.000	0.022	0.351	0.056
	4	0.000	0.000	0.003	0.067

		Delta Plot Purified Classification			
		1	2	3	4
True Classification	1	0.199	0.028	0.000	0.000
	2	0.041	0.175	0.058	0.000
	3	0.000	0.027	0.351	0.051
	4	0.000	0.000	0.003	0.066

		b Plot Purified Classification			
		1	2	3	4
True Classification	1	0.200	0.028	0.000	0.000
	2	0.042	0.174	0.058	0.000
	3	0.000	0.027	0.350	0.051
	4	0.000	0.000	0.003	0.066

		RPU Purified Classification			
		1	2	3	4
True Classification	1	0.199	0.028	0.000	0.000
	2	0.041	0.175	0.058	0.000
	3	0.000	0.027	0.351	0.051
	4	0.000	0.000	0.003	0.066

Continued, next page.

Table D.1., cont’d.:

Performance Level Contingency Tables, True vs. Aberrant & Purified Classification, 5 Spread Items with -0.7 *a* & -0.5 *b* Parameter Drift, Administration 1 Correlated Thetas of 0.8, Administration 2 Normal (No Shift) Distribution

		Aberrant Classification			
		1	2	3	4
True Classification	1	0.190	0.037	0.000	0.000
	2	0.031	0.178	0.065	0.000
	3	0.000	0.024	0.362	0.043
	4	0.000	0.000	0.005	0.065

		Delta Plot Purified Classification			
		1	2	3	4
True Classification	1	0.197	0.030	0.000	0.000
	2	0.039	0.180	0.055	0.000
	3	0.000	0.030	0.360	0.039
	4	0.000	0.000	0.005	0.064

		b Plot Purified Classification			
		1	2	3	4
True Classification	1	0.197	0.030	0.000	0.000
	2	0.039	0.178	0.057	0.000
	3	0.000	0.028	0.357	0.043
	4	0.000	0.000	0.004	0.065

		RPU Purified Classification			
		1	2	3	4
True Classification	1	0.197	0.030	0.000	0.000
	2	0.039	0.180	0.055	0.000
	3	0.000	0.030	0.360	0.039
	4	0.000	0.000	0.005	0.064

Continued, next page.

Table D.1., cont'd.:

Performance Level Contingency Tables, True vs. Aberrant & Purified Classification, 5 Spread Items with -0.7 *a* & -0.5 *b* Parameter Drift, Administration 1 Correlated Thetas of 0.8, Administration 2 Normal (No Shift) Distribution

		Aberrant Classification			
		1	2	3	4
True Classification	1	0.190	0.037	0.000	0.000
	2	0.031	0.178	0.064	0.000
	3	0.000	0.024	0.362	0.042
	4	0.000	0.000	0.005	0.065

		Delta Plot Purified Classification			
		1	2	3	4
True Classification	1	0.198	0.030	0.000	0.000
	2	0.039	0.180	0.054	0.000
	3	0.000	0.030	0.360	0.039
	4	0.000	0.000	0.005	0.064

		b Plot Purified Classification			
		1	2	3	4
True Classification	1	0.198	0.030	0.000	0.000
	2	0.040	0.178	0.056	0.000
	3	0.000	0.029	0.357	0.043
	4	0.000	0.000	0.005	0.065

		RPU Purified Classification			
		1	2	3	4
True Classification	1	0.197	0.030	0.000	0.000
	2	0.039	0.181	0.054	0.000
	3	0.000	0.030	0.360	0.038
	4	0.000	0.000	0.005	0.064

Continued, next page.

Table D.1., cont'd.:

Performance Level Contingency Tables, True vs. Aberrant & Purified Classification, 5 Spread Items with -0.7 *a* & -0.8 *b* Parameter Drift, Administration 1 Correlated Thetas of 0.8, Administration 2 Normal (No Shift) Distribution

		Aberrant Classification			
		1	2	3	4
True Classification	1	0.187	0.040	0.000	0.000
	2	0.028	0.177	0.069	0.000
	3	0.000	0.021	0.363	0.045
	4	0.000	0.000	0.004	0.065

		Delta Plot Purified Classification			
		1	2	3	4
True Classification	1	0.194	0.033	0.000	0.000
	2	0.035	0.180	0.059	0.000
	3	0.000	0.027	0.361	0.040
	4	0.000	0.000	0.005	0.065

		b Plot Purified Classification			
		1	2	3	4
True Classification	1	0.196	0.031	0.000	0.000
	2	0.037	0.177	0.060	0.000
	3	0.000	0.026	0.356	0.046
	4	0.000	0.000	0.004	0.066

		RPU Purified Classification			
		1	2	3	4
True Classification	1	0.197	0.030	0.000	0.000
	2	0.039	0.179	0.055	0.000
	3	0.000	0.029	0.359	0.040
	4	0.000	0.000	0.005	0.065

Continued, next page.

Table D.1., cont'd.:

Performance Level Contingency Tables, True vs. Aberrant & Purified Classification, 1 Hard Item with -0.3 *a* & -0.5 *b* Parameter Drift, Administration 1 Correlated Thetas of 0.8, Administration 2 Negatively Skewed (Shift) Distribution

		Aberrant Classification			
		1	2	3	4
True Classification	1	0.090	0.015	0.000	0.000
	2	0.032	0.207	0.097	0.000
	3	0.000	0.027	0.376	0.063
	4	0.000	0.000	0.004	0.087

		Delta Plot Purified Classification			
		1	2	3	4
True Classification	1	0.090	0.015	0.000	0.000
	2	0.033	0.209	0.096	0.000
	3	0.000	0.028	0.377	0.061
	4	0.000	0.000	0.004	0.087

		b Plot Purified Classification			
		1	2	3	4
True Classification	1	0.090	0.015	0.000	0.000
	2	0.033	0.209	0.095	0.000
	3	0.000	0.028	0.378	0.060
	4	0.000	0.000	0.004	0.087

		RPU Purified Classification			
		1	2	3	4
True Classification	1	0.094	0.011	0.000	0.000
	2	0.044	0.217	0.076	0.000
	3	0.000	0.037	0.376	0.053
	4	0.000	0.000	0.005	0.086

Continued, next page.

Table D.1., cont'd.:

Performance Level Contingency Tables, True vs. Aberrant & Purified Classification, 1 Hard Item with -0.7 *a* & -0.5 *b* Parameter Drift, Administration 1 Correlated Thetas of 0.8, Administration 2 Negatively Skewed (Shift) Distribution

		Aberrant Classification			
		1	2	3	4
True Classification	1	0.090	0.015	0.000	0.000
	2	0.033	0.208	0.097	0.000
	3	0.000	0.027	0.376	0.063
	4	0.000	0.000	0.004	0.087

		Delta Plot Purified Classification			
		1	2	3	4
True Classification	1	0.090	0.015	0.000	0.000
	2	0.033	0.209	0.096	0.000
	3	0.000	0.028	0.377	0.061
	4	0.000	0.000	0.004	0.087

		b Plot Purified Classification			
		1	2	3	4
True Classification	1	0.090	0.015	0.000	0.000
	2	0.033	0.209	0.095	0.000
	3	0.000	0.028	0.379	0.060
	4	0.000	0.000	0.004	0.087

		RPU Purified Classification			
		1	2	3	4
True Classification	1	0.094	0.011	0.000	0.000
	2	0.044	0.216	0.077	0.000
	3	0.000	0.036	0.375	0.055
	4	0.000	0.000	0.005	0.086

Continued, next page.

Table D.1., cont'd.:

Performance Level Contingency Tables, True vs. Aberrant & Purified Classification, 1 Hard Item with -0.3 *a* & -0.8 *b* Parameter Drift, Administration 1 Correlated Thetas of 0.8, Administration 2 Negatively Skewed (Shift) Distribution

		Aberrant Classification			
		1	2	3	4
True Classification	1	0.089	0.016	0.000	0.000
	2	0.032	0.206	0.099	0.000
	3	0.000	0.027	0.375	0.064
	4	0.000	0.000	0.004	0.088

		Delta Plot Purified Classification			
		1	2	3	4
True Classification	1	0.090	0.015	0.000	0.000
	2	0.033	0.209	0.096	0.000
	3	0.000	0.028	0.377	0.061
	4	0.000	0.000	0.004	0.087

		b Plot Purified Classification			
		1	2	3	4
True Classification	1	0.090	0.015	0.000	0.000
	2	0.033	0.209	0.096	0.000
	3	0.000	0.028	0.378	0.061
	4	0.000	0.000	0.004	0.087

		RPU Purified Classification			
		1	2	3	4
True Classification	1	0.093	0.012	0.000	0.000
	2	0.043	0.216	0.078	0.000
	3	0.000	0.036	0.376	0.054
	4	0.000	0.000	0.005	0.086

Continued, next page.

Table D.1., cont'd.:

Performance Level Contingency Tables, True vs. Aberrant & Purified Classification, 1 Hard Item with -0.7 *a* & -0.8 *b* Parameter Drift, Administration 1 Correlated Thetas of 0.8, Administration 2 Negatively Skewed (Shift) Distribution

		Aberrant Classification			
		1	2	3	4
True Classification	1	0.090	0.015	0.000	0.000
	2	0.032	0.207	0.098	0.000
	3	0.000	0.027	0.376	0.064
	4	0.000	0.000	0.004	0.088

		Delta Plot Purified Classification			
		1	2	3	4
True Classification	1	0.090	0.015	0.000	0.000
	2	0.033	0.209	0.095	0.000
	3	0.000	0.028	0.378	0.061
	4	0.000	0.000	0.004	0.087

		b Plot Purified Classification			
		1	2	3	4
True Classification	1	0.090	0.015	0.000	0.000
	2	0.033	0.209	0.095	0.000
	3	0.000	0.028	0.378	0.060
	4	0.000	0.000	0.004	0.087

		RPU Purified Classification			
		1	2	3	4
True Classification	1	0.093	0.011	0.000	0.000
	2	0.043	0.215	0.079	0.000
	3	0.000	0.036	0.375	0.056
	4	0.000	0.000	0.005	0.086

Continued, next page.

Table D.1., cont'd.:

Performance Level Contingency Tables, True vs. Aberrant & Purified Classification, 3 Spread Items with -0.3 *a* & -0.5 *b* Parameter Drift, Administration 1 Correlated Thetas of 0.8, Administration 2 Negatively Skewed (Shift) Distribution

		Aberrant Classification			
		1	2	3	4
True Classification	1	0.088	0.017	0.000	0.000
	2	0.028	0.205	0.103	0.000
	3	0.000	0.025	0.380	0.062
	4	0.000	0.000	0.004	0.087

		Delta Plot Purified Classification			
		1	2	3	4
True Classification	1	0.088	0.017	0.000	0.000
	2	0.029	0.207	0.102	0.000
	3	0.000	0.026	0.381	0.060
	4	0.000	0.000	0.004	0.087

		b Plot Purified Classification			
		1	2	3	4
True Classification	1	0.088	0.017	0.000	0.000
	2	0.029	0.207	0.101	0.000
	3	0.000	0.026	0.382	0.059
	4	0.000	0.000	0.004	0.087

		RPU Purified Classification			
		1	2	3	4
True Classification	1	0.093	0.012	0.000	0.000
	2	0.043	0.217	0.077	0.000
	3	0.000	0.037	0.378	0.051
	4	0.000	0.000	0.006	0.086

Continued, next page.

Table D.1., cont'd.:

Performance Level Contingency Tables, True vs. Aberrant & Purified Classification, 3 Spread Items with -0.7 *a* & -0.5 *b* Parameter Drift, Administration 1 Correlated Thetas of 0.8, Administration 2 Negatively Skewed (Shift) Distribution

		Aberrant Classification			
		1	2	3	4
True Classification	1	0.088	0.017	0.000	0.000
	2	0.029	0.206	0.103	0.000
	3	0.000	0.025	0.380	0.061
	4	0.000	0.000	0.004	0.087

		Delta Plot Purified Classification			
		1	2	3	4
True Classification	1	0.088	0.017	0.000	0.000
	2	0.029	0.207	0.102	0.000
	3	0.000	0.025	0.381	0.060
	4	0.000	0.000	0.004	0.087

		b Plot Purified Classification			
		1	2	3	4
True Classification	1	0.088	0.017	0.000	0.000
	2	0.029	0.207	0.101	0.000
	3	0.000	0.026	0.382	0.058
	4	0.000	0.000	0.004	0.087

		RPU Purified Classification			
		1	2	3	4
True Classification	1	0.093	0.012	0.000	0.000
	2	0.042	0.216	0.079	0.000
	3	0.000	0.036	0.377	0.053
	4	0.000	0.000	0.005	0.086

Continued, next page.

Table D.1., cont’d.:

Performance Level Contingency Tables, True vs. Aberrant & Purified Classification, 3 Spread Items with -0.7 *a* & -0.5 *b* Parameter Drift, Administration 1 Correlated Thetas of 0.8, Administration 2 Negatively Skewed (Shift) Distribution

		Aberrant Classification			
		1	2	3	4
True Classification	1	0.086	0.018	0.000	0.000
	2	0.026	0.202	0.109	0.000
	3	0.000	0.023	0.379	0.064
	4	0.000	0.000	0.004	0.088

		Delta Plot Purified Classification			
		1	2	3	4
True Classification	1	0.088	0.017	0.000	0.000
	2	0.029	0.207	0.102	0.000
	3	0.000	0.026	0.382	0.059
	4	0.000	0.000	0.004	0.087

		b Plot Purified Classification			
		1	2	3	4
True Classification	1	0.088	0.017	0.000	0.000
	2	0.029	0.207	0.101	0.000
	3	0.000	0.026	0.381	0.059
	4	0.000	0.000	0.004	0.087

		RPU Purified Classification			
		1	2	3	4
True Classification	1	0.093	0.012	0.000	0.000
	2	0.042	0.217	0.078	0.000
	3	0.000	0.036	0.378	0.052
	4	0.000	0.000	0.005	0.086

Continued, next page.

Table D.1., cont'd.:

Performance Level Contingency Tables, True vs. Aberrant & Purified Classification, 3 Spread Items with -0.7 *a* & -0.8 *b* Parameter Drift, Administration 1 Correlated Thetas of 0.8, Administration 2 Negatively Skewed (Shift) Distribution

		Aberrant Classification			
		1	2	3	4
True Classification	1	0.087	0.018	0.000	0.000
	2	0.026	0.202	0.109	0.000
	3	0.000	0.023	0.379	0.064
	4	0.000	0.000	0.004	0.088

		Delta Plot Purified Classification			
		1	2	3	4
True Classification	1	0.088	0.017	0.000	0.000
	2	0.029	0.207	0.102	0.000
	3	0.000	0.026	0.382	0.059
	4	0.000	0.000	0.004	0.087

		b Plot Purified Classification			
		1	2	3	4
True Classification	1	0.087	0.017	0.000	0.000
	2	0.028	0.206	0.103	0.000
	3	0.000	0.025	0.382	0.059
	4	0.000	0.000	0.004	0.087

		RPU Purified Classification			
		1	2	3	4
True Classification	1	0.093	0.012	0.000	0.000
	2	0.042	0.216	0.079	0.000
	3	0.000	0.036	0.377	0.054
	4	0.000	0.000	0.005	0.086

Continued, next page.

Table D.1., cont'd.:

Performance Level Contingency Tables, True vs. Aberrant & Purified Classification, 3 Moderate Items with -0.3 *a* & -0.5 *b* Parameter Drift, Administration 1 Correlated Thetas of 0.8, Administration 2 Negatively Skewed (Shift) Distribution

		Aberrant Classification			
		1	2	3	4
True Classification	1	0.087	0.018	0.000	0.000
	2	0.028	0.203	0.107	0.000
	3	0.000	0.024	0.378	0.065
	4	0.000	0.000	0.004	0.088

		Delta Plot Purified Classification			
		1	2	3	4
True Classification	1	0.087	0.018	0.000	0.000
	2	0.028	0.204	0.105	0.000
	3	0.000	0.024	0.379	0.063
	4	0.000	0.000	0.004	0.087

		b Plot Purified Classification			
		1	2	3	4
True Classification	1	0.087	0.017	0.000	0.000
	2	0.028	0.205	0.104	0.000
	3	0.000	0.025	0.380	0.062
	4	0.000	0.000	0.004	0.087

		RPU Purified Classification			
		1	2	3	4
True Classification	1	0.092	0.013	0.000	0.000
	2	0.039	0.216	0.082	0.000
	3	0.000	0.034	0.380	0.052
	4	0.000	0.000	0.005	0.086

Continued, next page.

Table D.1., cont'd.:

Performance Level Contingency Tables, True vs. Aberrant & Purified Classification, 3 Moderate Items with -0.7 *a* & -0.5 *b* Parameter Drift, Administration 1 Correlated Thetas of 0.8, Administration 2 Negatively Skewed (Shift) Distribution

		Aberrant Classification			
		1	2	3	4
True Classification	1	0.082	0.023	0.000	0.000
	2	0.020	0.196	0.122	0.000
	3	0.000	0.019	0.385	0.062
	4	0.000	0.000	0.004	0.087

		Delta Plot Purified Classification			
		1	2	3	4
True Classification	1	0.086	0.019	0.000	0.000
	2	0.026	0.206	0.105	0.000
	3	0.000	0.025	0.386	0.056
	4	0.000	0.000	0.005	0.086

		b Plot Purified Classification			
		1	2	3	4
True Classification	1	0.087	0.018	0.000	0.000
	2	0.027	0.207	0.104	0.000
	3	0.000	0.025	0.385	0.057
	4	0.000	0.000	0.005	0.086

		RPU Purified Classification			
		1	2	3	4
True Classification	1	0.091	0.014	0.000	0.000
	2	0.035	0.219	0.083	0.000
	3	0.000	0.034	0.387	0.045
	4	0.000	0.000	0.007	0.084

Continued, next page.

Table D.1., cont'd.:

Performance Level Contingency Tables, True vs. Aberrant & Purified Classification, 3 Moderate Items with -0.7 *a* & -0.5 *b* Parameter Drift, Administration 1 Correlated Thetas of 0.8, Administration 2 Negatively Skewed (Shift) Distribution

		Aberrant Classification			
		1	2	3	4
True Classification	1	0.086	0.019	0.000	0.000
	2	0.026	0.199	0.112	0.000
	3	0.000	0.022	0.376	0.068
	4	0.000	0.000	0.003	0.088

		Delta Plot Purified Classification			
		1	2	3	4
True Classification	1	0.087	0.017	0.000	0.000
	2	0.028	0.205	0.104	0.000
	3	0.000	0.025	0.379	0.063
	4	0.000	0.000	0.004	0.087

		b Plot Purified Classification			
		1	2	3	4
True Classification	1	0.088	0.017	0.000	0.000
	2	0.028	0.205	0.104	0.000
	3	0.000	0.025	0.380	0.062
	4	0.000	0.000	0.004	0.087

		RPU Purified Classification			
		1	2	3	4
True Classification	1	0.092	0.013	0.000	0.000
	2	0.040	0.217	0.081	0.000
	3	0.000	0.035	0.380	0.052
	4	0.000	0.000	0.006	0.086

Continued, next page.

Table D.1., cont’d.:

Performance Level Contingency Tables, True vs. Aberrant & Purified Classification, 3 Moderate Items with -0.7 *a* & -0.8 *b* Parameter Drift, Administration 1 Correlated Thetas of 0.8, Administration 2 Negatively Skewed (Shift) Distribution

		Aberrant Classification			
		1	2	3	4
True Classification	1	0.082	0.023	0.000	0.000
	2	0.019	0.195	0.123	0.000
	3	0.000	0.019	0.385	0.063
	4	0.000	0.000	0.004	0.087

		Delta Plot Purified Classification			
		1	2	3	4
True Classification	1	0.086	0.019	0.000	0.000
	2	0.025	0.205	0.107	0.000
	3	0.000	0.024	0.386	0.057
	4	0.000	0.000	0.005	0.086

		b Plot Purified Classification			
		1	2	3	4
True Classification	1	0.086	0.019	0.000	0.000
	2	0.026	0.206	0.106	0.000
	3	0.000	0.024	0.385	0.057
	4	0.000	0.000	0.005	0.087

		RPU Purified Classification			
		1	2	3	4
True Classification	1	0.092	0.013	0.000	0.000
	2	0.039	0.219	0.079	0.000
	3	0.000	0.036	0.383	0.047
	4	0.000	0.000	0.007	0.085

Continued, next page.

Table D.1., cont'd.:

Performance Level Contingency Tables, True vs. Aberrant & Purified Classification, 5 Spread Items with -0.3 *a* & -0.5 *b* Parameter Drift, Administration 1 Correlated Thetas of 0.8, Administration 2 Negatively Skewed (Shift) Distribution

		Aberrant Classification			
		1	2	3	4
True Classification	1	0.085	0.020	0.000	0.000
	2	0.024	0.202	0.111	0.000
	3	0.000	0.022	0.382	0.062
	4	0.000	0.000	0.004	0.087

		Delta Plot Purified Classification			
		1	2	3	4
True Classification	1	0.086	0.019	0.000	0.000
	2	0.025	0.204	0.108	0.000
	3	0.000	0.023	0.383	0.060
	4	0.000	0.000	0.004	0.087

		b Plot Purified Classification			
		1	2	3	4
True Classification	1	0.085	0.020	0.000	0.000
	2	0.025	0.204	0.109	0.000
	3	0.000	0.023	0.384	0.059
	4	0.000	0.000	0.004	0.087

		RPU Purified Classification			
		1	2	3	4
True Classification	1	0.092	0.013	0.000	0.000
	2	0.039	0.216	0.081	0.000
	3	0.000	0.035	0.380	0.052
	4	0.000	0.000	0.006	0.086

Continued, next page.

Table D.1., cont'd.:

Performance Level Contingency Tables, True vs. Aberrant & Purified Classification, 5 Spread Items with -0.7 *a* & -0.5 *b* Parameter Drift, Administration 1 Correlated Thetas of 0.8, Administration 2 Negatively Skewed (Shift) Distribution

		Aberrant Classification			
		1	2	3	4
True Classification	1	0.080	0.025	0.000	0.000
	2	0.018	0.202	0.118	0.000
	3	0.000	0.021	0.398	0.048
	4	0.000	0.000	0.006	0.085

		Delta Plot Purified Classification			
		1	2	3	4
True Classification	1	0.085	0.019	0.000	0.000
	2	0.025	0.210	0.102	0.000
	3	0.000	0.026	0.392	0.049
	4	0.000	0.000	0.006	0.085

		b Plot Purified Classification			
		1	2	3	4
True Classification	1	0.085	0.020	0.000	0.000
	2	0.025	0.210	0.103	0.000
	3	0.000	0.026	0.392	0.048
	4	0.000	0.000	0.006	0.085

		RPU Purified Classification			
		1	2	3	4
True Classification	1	0.089	0.016	0.000	0.000
	2	0.032	0.224	0.081	0.000
	3	0.000	0.035	0.396	0.036
	4	0.000	0.000	0.010	0.081

Continued, next page.

Table D.1., cont'd.:

Performance Level Contingency Tables, True vs. Aberrant & Purified Classification, 5 Spread Items with -0.7 *a* & -0.5 *b* Parameter Drift, Administration 1 Correlated Thetas of 0.8, Administration 2 Negatively Skewed (Shift) Distribution

		Aberrant Classification			
		1	2	3	4
True Classification	1	0.083	0.022	0.000	0.000
	2	0.021	0.195	0.122	0.000
	3	0.000	0.019	0.382	0.066
	4	0.000	0.000	0.004	0.088

		Delta Plot Purified Classification			
		1	2	3	4
True Classification	1	0.085	0.020	0.000	0.000
	2	0.024	0.202	0.111	0.000
	3	0.000	0.022	0.382	0.062
	4	0.000	0.000	0.004	0.087

		b Plot Purified Classification			
		1	2	3	4
True Classification	1	0.085	0.020	0.000	0.000
	2	0.024	0.202	0.112	0.000
	3	0.000	0.022	0.384	0.060
	4	0.000	0.000	0.004	0.087

		RPU Purified Classification			
		1	2	3	4
True Classification	1	0.092	0.013	0.000	0.000
	2	0.040	0.217	0.080	0.000
	3	0.000	0.035	0.380	0.051
	4	0.000	0.000	0.006	0.086

Continued, next page.

Table D.1., cont'd.:

Performance Level Contingency Tables, True vs. Aberrant & Purified Classification, 5 Spread Items with -0.7 *a* & -0.8 *b* Parameter Drift, Administration 1 Correlated Thetas of 0.8, Administration 2 Negatively Skewed (Shift) Distribution

		Aberrant Classification			
		1	2	3	4
True Classification	1	0.078	0.026	0.000	0.000
	2	0.016	0.197	0.124	0.000
	3	0.000	0.019	0.398	0.049
	4	0.000	0.000	0.006	0.085

		Delta Plot Purified Classification			
		1	2	3	4
True Classification	1	0.084	0.021	0.000	0.000
	2	0.023	0.206	0.109	0.000
	3	0.000	0.024	0.392	0.051
	4	0.000	0.000	0.006	0.085

		b Plot Purified Classification			
		1	2	3	4
True Classification	1	0.084	0.021	0.000	0.000
	2	0.023	0.207	0.108	0.000
	3	0.000	0.024	0.391	0.051
	4	0.000	0.000	0.006	0.085

		RPU Purified Classification			
		1	2	3	4
True Classification	1	0.091	0.014	0.000	0.000
	2	0.036	0.225	0.076	0.000
	3	0.000	0.038	0.392	0.036
	4	0.000	0.000	0.010	0.082

Continued, next page.

Table D.1., cont'd.:

Performance Level Contingency Tables, True vs. Aberrant & Purified Classification, 1 Hard Item with -0.3 *a* & -0.5 *b* Parameter Drift, Administration 1 Correlated Thetas of 0.6, Administration 2 Normal (No Shift) Distribution

		Aberrant Classification			
		1	2	3	4
True Classification	1	0.204	0.023	0.000	0.000
	2	0.049	0.172	0.053	0.000
	3	0.000	0.030	0.345	0.053
	4	0.000	0.000	0.003	0.067

		Delta Plot Purified Classification			
		1	2	3	4
True Classification	1	0.205	0.023	0.000	0.000
	2	0.049	0.173	0.052	0.000
	3	0.000	0.031	0.346	0.052
	4	0.000	0.000	0.003	0.067

		b Plot Purified Classification			
		1	2	3	4
True Classification	1	0.204	0.024	0.000	0.000
	2	0.048	0.173	0.053	0.000
	3	0.000	0.030	0.347	0.051
	4	0.000	0.000	0.003	0.067

		RPU Purified Classification			
		1	2	3	4
True Classification	1	0.204	0.023	0.000	0.000
	2	0.049	0.172	0.053	0.000
	3	0.000	0.030	0.345	0.053
	4	0.000	0.000	0.003	0.067

Continued, next page.

Table D.1., cont’d.:

Performance Level Contingency Tables, True vs. Aberrant & Purified Classification, 1 Hard Item with -0.7 *a* & -0.5 *b* Parameter Drift, Administration 1 Correlated Thetas of 0.6, Administration 2 Normal (No Shift) Distribution

		Aberrant Classification			
		1	2	3	4
True Classification	1	0.198	0.030	0.000	0.000
	2	0.039	0.172	0.064	0.000
	3	0.000	0.024	0.347	0.058
	4	0.000	0.000	0.003	0.067

		Delta Plot Purified Classification			
		1	2	3	4
True Classification	1	0.204	0.023	0.000	0.000
	2	0.049	0.172	0.052	0.000
	3	0.000	0.031	0.346	0.052
	4	0.000	0.000	0.003	0.067

		b Plot Purified Classification			
		1	2	3	4
True Classification	1	0.204	0.023	0.000	0.000
	2	0.049	0.173	0.052	0.000
	3	0.000	0.031	0.346	0.052
	4	0.000	0.000	0.003	0.067

		RPU Purified Classification			
		1	2	3	4
True Classification	1	0.204	0.023	0.000	0.000
	2	0.049	0.172	0.052	0.000
	3	0.000	0.031	0.346	0.052
	4	0.000	0.000	0.003	0.067

Continued, next page.

Table D.1., cont'd.:

Performance Level Contingency Tables, True vs. Aberrant & Purified Classification, 1 Hard Item with -0.3 *a* & -0.8 *b* Parameter Drift, Administration 1 Correlated Thetas of 0.6, Administration 2 Normal (No Shift) Distribution

		Aberrant Classification			
		1	2	3	4
True Classification	1	0.204	0.024	0.000	0.000
	2	0.048	0.172	0.054	0.000
	3	0.000	0.029	0.345	0.054
	4	0.000	0.000	0.003	0.067

		Delta Plot Purified Classification			
		1	2	3	4
True Classification	1	0.205	0.023	0.000	0.000
	2	0.049	0.173	0.052	0.000
	3	0.000	0.031	0.346	0.052
	4	0.000	0.000	0.003	0.067

		b Plot Purified Classification			
		1	2	3	4
True Classification	1	0.204	0.023	0.000	0.000
	2	0.049	0.173	0.052	0.000
	3	0.000	0.031	0.347	0.051
	4	0.000	0.000	0.003	0.067

		RPU Purified Classification			
		1	2	3	4
True Classification	1	0.204	0.024	0.000	0.000
	2	0.048	0.172	0.054	0.000
	3	0.000	0.029	0.345	0.054
	4	0.000	0.000	0.003	0.067

Continued, next page.

Table D.1., cont'd.:

Performance Level Contingency Tables, True vs. Aberrant & Purified Classification, 1 Hard Item with -0.7 *a* & -0.8 *b* Parameter Drift, Administration 1 Correlated Thetas of 0.6, Administration 2 Normal (No Shift) Distribution

		Aberrant Classification			
		1	2	3	4
True Classification	1	0.198	0.030	0.000	0.000
	2	0.039	0.172	0.064	0.000
	3	0.000	0.024	0.347	0.058
	4	0.000	0.000	0.003	0.067

		Delta Plot Purified Classification			
		1	2	3	4
True Classification	1	0.204	0.023	0.000	0.000
	2	0.049	0.172	0.052	0.000
	3	0.000	0.031	0.346	0.052
	4	0.000	0.000	0.003	0.067

		b Plot Purified Classification			
		1	2	3	4
True Classification	1	0.204	0.023	0.000	0.000
	2	0.049	0.173	0.052	0.000
	3	0.000	0.031	0.346	0.052
	4	0.000	0.000	0.003	0.067

		RPU Purified Classification			
		1	2	3	4
True Classification	1	0.204	0.023	0.000	0.000
	2	0.049	0.172	0.052	0.000
	3	0.000	0.031	0.346	0.052
	4	0.000	0.000	0.003	0.067

Continued, next page.

Table D.1., cont'd.:

Performance Level Contingency Tables, True vs. Aberrant & Purified Classification, 3 Spread Items with -0.3 *a* & -0.5 *b* Parameter Drift, Administration 1 Correlated Thetas of 0.6, Administration 2 Normal (No Shift) Distribution

		Aberrant Classification			
		1	2	3	4
True Classification	1	0.201	0.026	0.000	0.000
	2	0.044	0.174	0.057	0.000
	3	0.000	0.028	0.349	0.052
	4	0.000	0.000	0.003	0.067

		Delta Plot Purified Classification			
		1	2	3	4
True Classification	1	0.201	0.026	0.000	0.000
	2	0.044	0.174	0.056	0.000
	3	0.000	0.028	0.349	0.051
	4	0.000	0.000	0.003	0.067

		b Plot Purified Classification			
		1	2	3	4
True Classification	1	0.201	0.027	0.000	0.000
	2	0.043	0.175	0.056	0.000
	3	0.000	0.028	0.351	0.050
	4	0.000	0.000	0.003	0.066

		RPU Purified Classification			
		1	2	3	4
True Classification	1	0.201	0.026	0.000	0.000
	2	0.044	0.174	0.057	0.000
	3	0.000	0.028	0.349	0.052
	4	0.000	0.000	0.003	0.067

Continued, next page.

Table D.1., cont'd.:

Performance Level Contingency Tables, True vs. Aberrant & Purified Classification, 3 Spread Items with -0.7 *a* & -0.5 *b* Parameter Drift, Administration 1 Correlated Thetas of 0.6, Administration 2 Normal (No Shift) Distribution

		Aberrant Classification			
		1	2	3	4
True Classification	1	0.194	0.033	0.000	0.000
	2	0.035	0.177	0.062	0.000
	3	0.000	0.025	0.357	0.047
	4	0.000	0.000	0.004	0.066

		Delta Plot Purified Classification			
		1	2	3	4
True Classification	1	0.201	0.027	0.000	0.000
	2	0.044	0.178	0.053	0.000
	3	0.000	0.031	0.355	0.043
	4	0.000	0.000	0.005	0.065

		b Plot Purified Classification			
		1	2	3	4
True Classification	1	0.201	0.027	0.000	0.000
	2	0.044	0.176	0.055	0.000
	3	0.000	0.029	0.352	0.048
	4	0.000	0.000	0.004	0.066

		RPU Purified Classification			
		1	2	3	4
True Classification	1	0.201	0.026	0.000	0.000
	2	0.044	0.178	0.052	0.000
	3	0.000	0.031	0.354	0.043
	4	0.000	0.000	0.004	0.065

Continued, next page.

Table D.1., cont'd.:

Performance Level Contingency Tables, True vs. Aberrant & Purified Classification, 3 Spread Items with -0.7 *a* & -0.5 *b* Parameter Drift, Administration 1 Correlated Thetas of 0.6, Administration 2 Normal (No Shift) Distribution

		Aberrant Classification			
		1	2	3	4
True Classification	1	0.199	0.028	0.000	0.000
	2	0.041	0.173	0.060	0.000
	3	0.000	0.026	0.349	0.054
	4	0.000	0.000	0.003	0.067

		Delta Plot Purified Classification			
		1	2	3	4
True Classification	1	0.204	0.024	0.000	0.000
	2	0.048	0.173	0.052	0.000
	3	0.000	0.031	0.347	0.051
	4	0.000	0.000	0.003	0.067

		b Plot Purified Classification			
		1	2	3	4
True Classification	1	0.201	0.026	0.000	0.000
	2	0.044	0.174	0.056	0.000
	3	0.000	0.028	0.349	0.051
	4	0.000	0.000	0.003	0.066

		RPU Purified Classification			
		1	2	3	4
True Classification	1	0.203	0.024	0.000	0.000
	2	0.047	0.173	0.054	0.000
	3	0.000	0.030	0.346	0.053
	4	0.000	0.000	0.003	0.067

Continued, next page.

Table D.1., cont'd.:

Performance Level Contingency Tables, True vs. Aberrant & Purified Classification, 3 Spread Items with -0.7 *a* & -0.8 *b* Parameter Drift, Administration 1 Correlated Thetas of 0.6, Administration 2 Normal (No Shift) Distribution

		Aberrant Classification			
		1	2	3	4
True Classification	1	0.193	0.034	0.000	0.000
	2	0.034	0.177	0.064	0.000
	3	0.000	0.024	0.358	0.047
	4	0.000	0.000	0.004	0.066

		Delta Plot Purified Classification			
		1	2	3	4
True Classification	1	0.200	0.028	0.000	0.000
	2	0.042	0.178	0.054	0.000
	3	0.000	0.030	0.356	0.043
	4	0.000	0.000	0.004	0.065

		b Plot Purified Classification			
		1	2	3	4
True Classification	1	0.200	0.027	0.000	0.000
	2	0.042	0.175	0.056	0.000
	3	0.000	0.028	0.352	0.048
	4	0.000	0.000	0.004	0.066

		RPU Purified Classification			
		1	2	3	4
True Classification	1	0.200	0.027	0.000	0.000
	2	0.043	0.178	0.053	0.000
	3	0.000	0.030	0.355	0.043
	4	0.000	0.000	0.005	0.065

Continued, next page.

Table D.1., cont'd.:

Performance Level Contingency Tables, True vs. Aberrant & Purified Classification, 3 Moderate Items with -0.3 *a* & -0.5 *b* Parameter Drift, Administration 1 Correlated Thetas of 0.6, Administration 2 Normal (No Shift) Distribution

		Aberrant Classification			
		1	2	3	4
True Classification	1	0.194	0.034	0.000	0.000
	2	0.034	0.177	0.063	0.000
	3	0.000	0.024	0.358	0.046
	4	0.000	0.000	0.004	0.066

		Delta Plot Purified Classification			
		1	2	3	4
True Classification	1	0.200	0.027	0.000	0.000
	2	0.043	0.178	0.053	0.000
	3	0.000	0.030	0.355	0.043
	4	0.000	0.000	0.005	0.065

		b Plot Purified Classification			
		1	2	3	4
True Classification	1	0.200	0.027	0.000	0.000
	2	0.043	0.175	0.056	0.000
	3	0.000	0.029	0.351	0.048
	4	0.000	0.000	0.004	0.066

		RPU Purified Classification			
		1	2	3	4
True Classification	1	0.200	0.027	0.000	0.000
	2	0.043	0.178	0.053	0.000
	3	0.000	0.031	0.355	0.042
	4	0.000	0.000	0.005	0.065

Continued, next page.

Table D.1., cont'd.:

Performance Level Contingency Tables, True vs. Aberrant & Purified Classification, 3 Moderate Items with -0.7 *a* & -0.5 *b* Parameter Drift, Administration 1 Correlated Thetas of 0.6, Administration 2 Normal (No Shift) Distribution

		Aberrant Classification			
		1	2	3	4
True Classification	1	0.192	0.035	0.000	0.000
	2	0.032	0.174	0.068	0.000
	3	0.000	0.022	0.355	0.052
	4	0.000	0.000	0.003	0.067

		Delta Plot Purified Classification			
		1	2	3	4
True Classification	1	0.199	0.028	0.000	0.000
	2	0.041	0.176	0.058	0.000
	3	0.000	0.027	0.353	0.048
	4	0.000	0.000	0.004	0.066

		b Plot Purified Classification			
		1	2	3	4
True Classification	1	0.200	0.028	0.000	0.000
	2	0.042	0.175	0.057	0.000
	3	0.000	0.028	0.352	0.048
	4	0.000	0.000	0.004	0.066

		RPU Purified Classification			
		1	2	3	4
True Classification	1	0.199	0.028	0.000	0.000
	2	0.041	0.176	0.058	0.000
	3	0.000	0.027	0.353	0.048
	4	0.000	0.000	0.004	0.066

Continued, next page.

Table D.1., cont'd.:

Performance Level Contingency Tables, True vs. Aberrant & Purified Classification, 3 Moderate Items with -0.7 *a* & -0.5 *b* Parameter Drift, Administration 1 Correlated Thetas of 0.6, Administration 2 Normal (No Shift) Distribution

		Aberrant Classification			
		1	2	3	4
True Classification	1	0.192	0.035	0.000	0.000
	2	0.033	0.174	0.067	0.000
	3	0.000	0.022	0.355	0.052
	4	0.000	0.000	0.003	0.067

		Delta Plot Purified Classification			
		1	2	3	4
True Classification	1	0.199	0.028	0.000	0.000
	2	0.041	0.176	0.057	0.000
	3	0.000	0.028	0.353	0.048
	4	0.000	0.000	0.004	0.066

		b Plot Purified Classification			
		1	2	3	4
True Classification	1	0.200	0.028	0.000	0.000
	2	0.042	0.176	0.056	0.000
	3	0.000	0.028	0.352	0.048
	4	0.000	0.000	0.004	0.066

		RPU Purified Classification			
		1	2	3	4
True Classification	1	0.199	0.028	0.000	0.000
	2	0.041	0.176	0.057	0.000
	3	0.000	0.028	0.353	0.048
	4	0.000	0.000	0.004	0.066

Continued, next page.

Table D.1., cont’d.:

Performance Level Contingency Tables, True vs. Aberrant & Purified Classification, 3 Moderate Items with -0.7 *a* & -0.8 *b* Parameter Drift, Administration 1 Correlated Thetas of 0.6, Administration 2 Normal (No Shift) Distribution

		Aberrant Classification			
		1	2	3	4
True Classification	1	0.192	0.036	0.000	0.000
	2	0.032	0.174	0.069	0.000
	3	0.000	0.021	0.355	0.053
	4	0.000	0.000	0.003	0.067

		Delta Plot Purified Classification			
		1	2	3	4
True Classification	1	0.198	0.029	0.000	0.000
	2	0.040	0.176	0.059	0.000
	3	0.000	0.027	0.353	0.049
	4	0.000	0.000	0.004	0.066

		b Plot Purified Classification			
		1	2	3	4
True Classification	1	0.199	0.028	0.000	0.000
	2	0.041	0.175	0.058	0.000
	3	0.000	0.027	0.352	0.049
	4	0.000	0.000	0.004	0.066

		RPU Purified Classification			
		1	2	3	4
True Classification	1	0.198	0.029	0.000	0.000
	2	0.040	0.176	0.059	0.000
	3	0.000	0.027	0.353	0.049
	4	0.000	0.000	0.004	0.066

Continued, next page.

Table D.1., cont'd.:

Performance Level Contingency Tables, True vs. Aberrant & Purified Classification, 5 Spread Items with -0.3 *a* & -0.5 *b* Parameter Drift, Administration 1 Correlated Thetas of 0.6, Administration 2 Normal (No Shift) Distribution

		Aberrant Classification			
		1	2	3	4
True Classification	1	0.192	0.035	0.000	0.000
	2	0.032	0.174	0.068	0.000
	3	0.000	0.022	0.355	0.052
	4	0.000	0.000	0.003	0.067

		Delta Plot Purified Classification			
		1	2	3	4
True Classification	1	0.198	0.029	0.000	0.000
	2	0.040	0.176	0.058	0.000
	3	0.000	0.027	0.353	0.048
	4	0.000	0.000	0.004	0.066

		b Plot Purified Classification			
		1	2	3	4
True Classification	1	0.199	0.028	0.000	0.000
	2	0.041	0.175	0.057	0.000
	3	0.000	0.027	0.352	0.049
	4	0.000	0.000	0.004	0.066

		RPU Purified Classification			
		1	2	3	4
True Classification	1	0.199	0.029	0.000	0.000
	2	0.040	0.176	0.058	0.000
	3	0.000	0.027	0.353	0.048
	4	0.000	0.000	0.004	0.066

Continued, next page.

Table D.1., cont'd.:

Performance Level Contingency Tables, True vs. Aberrant & Purified Classification, 5 Spread Items with -0.7 *a* & -0.5 *b* Parameter Drift, Administration 1 Correlated Thetas of 0.6, Administration 2 Normal (No Shift) Distribution

		Aberrant Classification			
		1	2	3	4
True Classification	1	0.189	0.039	0.000	0.000
	2	0.029	0.179	0.065	0.000
	3	0.000	0.023	0.365	0.040
	4	0.000	0.000	0.005	0.065

		Delta Plot Purified Classification			
		1	2	3	4
True Classification	1	0.196	0.031	0.000	0.000
	2	0.037	0.181	0.055	0.000
	3	0.000	0.029	0.362	0.037
	4	0.000	0.000	0.006	0.064

		b Plot Purified Classification			
		1	2	3	4
True Classification	1	0.197	0.031	0.000	0.000
	2	0.038	0.179	0.057	0.000
	3	0.000	0.028	0.359	0.042
	4	0.000	0.000	0.005	0.065

		RPU Purified Classification			
		1	2	3	4
True Classification	1	0.196	0.031	0.000	0.000
	2	0.038	0.182	0.055	0.000
	3	0.000	0.030	0.362	0.037
	4	0.000	0.000	0.006	0.064

Continued, next page.

Table D.1., cont'd.:

Performance Level Contingency Tables, True vs. Aberrant & Purified Classification, 5 Spread Items with -0.7 *a* & -0.5 *b* Parameter Drift, Administration 1 Correlated Thetas of 0.6, Administration 2 Normal (No Shift) Distribution

		Aberrant Classification			
		1	2	3	4
True Classification	1	0.189	0.038	0.000	0.000
	2	0.030	0.180	0.065	0.000
	3	0.000	0.024	0.365	0.040
	4	0.000	0.000	0.005	0.065

		Delta Plot Purified Classification			
		1	2	3	4
True Classification	1	0.196	0.031	0.000	0.000
	2	0.038	0.182	0.054	0.000
	3	0.000	0.030	0.362	0.037
	4	0.000	0.000	0.006	0.064

		b Plot Purified Classification			
		1	2	3	4
True Classification	1	0.197	0.030	0.000	0.000
	2	0.039	0.179	0.056	0.000
	3	0.000	0.029	0.358	0.041
	4	0.000	0.000	0.005	0.065

		RPU Purified Classification			
		1	2	3	4
True Classification	1	0.196	0.031	0.000	0.000
	2	0.038	0.182	0.054	0.000
	3	0.000	0.030	0.362	0.036
	4	0.000	0.000	0.006	0.064

Continued, next page.

Table D.1., cont'd.:

Performance Level Contingency Tables, True vs. Aberrant & Purified Classification, 5 Spread Items with -0.7 *a* & -0.8 *b* Parameter Drift, Administration 1 Correlated Thetas of 0.6, Administration 2 Normal (No Shift) Distribution

		Aberrant Classification			
		1	2	3	4
True Classification	1	0.186	0.041	0.000	0.000
	2	0.027	0.178	0.069	0.000
	3	0.000	0.021	0.366	0.042
	4	0.000	0.000	0.005	0.065

		Delta Plot Purified Classification			
		1	2	3	4
True Classification	1	0.193	0.034	0.000	0.000
	2	0.034	0.181	0.059	0.000
	3	0.000	0.027	0.363	0.038
	4	0.000	0.000	0.006	0.064

		b Plot Purified Classification			
		1	2	3	4
True Classification	1	0.195	0.032	0.000	0.000
	2	0.036	0.178	0.060	0.000
	3	0.000	0.026	0.358	0.044
	4	0.000	0.000	0.004	0.065

		RPU Purified Classification			
		1	2	3	4
True Classification	1	0.197	0.031	0.000	0.000
	2	0.038	0.180	0.055	0.000
	3	0.000	0.029	0.361	0.039
	4	0.000	0.000	0.005	0.064

Continued, next page.

Table D.1., cont'd.:

Performance Level Contingency Tables, True vs. Aberrant & Purified Classification, 1 Hard Item with -0.3 *a* & -0.5 *b* Parameter Drift, Administration 1 Correlated Thetas of 0.6, Administration 2 Negatively Skewed (Shift) Distribution

		Aberrant Classification			
		1	2	3	4
True Classification	1	0.089	0.016	0.000	0.000
	2	0.031	0.209	0.098	0.000
	3	0.000	0.027	0.380	0.059
	4	0.000	0.000	0.004	0.087

		Delta Plot Purified Classification			
		1	2	3	4
True Classification	1	0.089	0.016	0.000	0.000
	2	0.032	0.210	0.096	0.000
	3	0.000	0.028	0.380	0.058
	4	0.000	0.000	0.004	0.087

		b Plot Purified Classification			
		1	2	3	4
True Classification	1	0.089	0.016	0.000	0.000
	2	0.031	0.211	0.095	0.000
	3	0.000	0.028	0.382	0.057
	4	0.000	0.000	0.005	0.087

		RPU Purified Classification			
		1	2	3	4
True Classification	1	0.093	0.012	0.000	0.000
	2	0.043	0.219	0.075	0.000
	3	0.000	0.037	0.379	0.049
	4	0.000	0.000	0.006	0.085

Continued, next page.

Table D.1., cont'd.:

Performance Level Contingency Tables, True vs. Aberrant & Purified Classification, 1 Hard Item with -0.7 *a* & -0.5 *b* Parameter Drift, Administration 1 Correlated Thetas of 0.6, Administration 2 Negatively Skewed (Shift) Distribution

		Aberrant Classification			
		1	2	3	4
True Classification	1	0.089	0.016	0.000	0.000
	2	0.031	0.209	0.097	0.000
	3	0.000	0.027	0.380	0.059
	4	0.000	0.000	0.004	0.087

		Delta Plot Purified Classification			
		1	2	3	4
True Classification	1	0.089	0.016	0.000	0.000
	2	0.032	0.210	0.096	0.000
	3	0.000	0.028	0.380	0.059
	4	0.000	0.000	0.004	0.087

		b Plot Purified Classification			
		1	2	3	4
True Classification	1	0.089	0.016	0.000	0.000
	2	0.031	0.210	0.096	0.000
	3	0.000	0.028	0.382	0.057
	4	0.000	0.000	0.005	0.087

		RPU Purified Classification			
		1	2	3	4
True Classification	1	0.093	0.012	0.000	0.000
	2	0.043	0.218	0.076	0.000
	3	0.000	0.037	0.378	0.051
	4	0.000	0.000	0.006	0.085

Continued, next page.

Table D.1., cont'd.:

Performance Level Contingency Tables, True vs. Aberrant & Purified Classification, 1 Hard Item with -0.3 *a* & -0.8 *b* Parameter Drift, Administration 1 Correlated Thetas of 0.6, Administration 2 Negatively Skewed (Shift) Distribution

		Aberrant Classification			
		1	2	3	4
True Classification	1	0.089	0.016	0.000	0.000
	2	0.030	0.208	0.099	0.000
	3	0.000	0.026	0.379	0.061
	4	0.000	0.000	0.004	0.087

		Delta Plot Purified Classification			
		1	2	3	4
True Classification	1	0.089	0.016	0.000	0.000
	2	0.032	0.210	0.096	0.000
	3	0.000	0.028	0.380	0.058
	4	0.000	0.000	0.004	0.087

		b Plot Purified Classification			
		1	2	3	4
True Classification	1	0.089	0.016	0.000	0.000
	2	0.031	0.211	0.095	0.000
	3	0.000	0.028	0.382	0.057
	4	0.000	0.000	0.005	0.087

		RPU Purified Classification			
		1	2	3	4
True Classification	1	0.093	0.012	0.000	0.000
	2	0.042	0.218	0.077	0.000
	3	0.000	0.037	0.379	0.051
	4	0.000	0.000	0.006	0.086

Continued, next page.

Table D.1., cont'd.:

Performance Level Contingency Tables, True vs. Aberrant & Purified Classification, 1 Hard Item with -0.7 *a* & -0.8 *b* Parameter Drift, Administration 1 Correlated Thetas of 0.6, Administration 2 Negatively Skewed (Shift) Distribution

		Aberrant Classification			
		1	2	3	4
True Classification	1	0.089	0.016	0.000	0.000
	2	0.031	0.208	0.099	0.000
	3	0.000	0.027	0.379	0.060
	4	0.000	0.000	0.004	0.087

		Delta Plot Purified Classification			
		1	2	3	4
True Classification	1	0.089	0.016	0.000	0.000
	2	0.032	0.210	0.096	0.000
	3	0.000	0.028	0.380	0.058
	4	0.000	0.000	0.005	0.087

		b Plot Purified Classification			
		1	2	3	4
True Classification	1	0.089	0.016	0.000	0.000
	2	0.031	0.210	0.096	0.000
	3	0.000	0.028	0.382	0.057
	4	0.000	0.000	0.005	0.087

		RPU Purified Classification			
		1	2	3	4
True Classification	1	0.093	0.012	0.000	0.000
	2	0.042	0.217	0.078	0.000
	3	0.000	0.036	0.378	0.052
	4	0.000	0.000	0.006	0.086

Continued, next page.

Table D.1., cont'd.:

Performance Level Contingency Tables, True vs. Aberrant & Purified Classification, 3 Spread Items with -0.3 *a* & -0.5 *b* Parameter Drift, Administration 1 Correlated Thetas of 0.6, Administration 2 Negatively Skewed (Shift) Distribution

		Aberrant Classification			
		1	2	3	4
True Classification	1	0.087	0.018	0.000	0.000
	2	0.027	0.207	0.104	0.000
	3	0.000	0.025	0.383	0.058
	4	0.000	0.000	0.005	0.087

		Delta Plot Purified Classification			
		1	2	3	4
True Classification	1	0.087	0.018	0.000	0.000
	2	0.028	0.208	0.102	0.000
	3	0.000	0.025	0.384	0.057
	4	0.000	0.000	0.005	0.087

		b Plot Purified Classification			
		1	2	3	4
True Classification	1	0.087	0.018	0.000	0.000
	2	0.027	0.208	0.102	0.000
	3	0.000	0.026	0.385	0.055
	4	0.000	0.000	0.005	0.086

		RPU Purified Classification			
		1	2	3	4
True Classification	1	0.093	0.012	0.000	0.000
	2	0.042	0.219	0.076	0.000
	3	0.000	0.037	0.380	0.049
	4	0.000	0.000	0.006	0.085

Continued, next page.

Table D.1., cont'd.:

Performance Level Contingency Tables, True vs. Aberrant & Purified Classification, 3 Spread Items with -0.7 *a* & -0.5 *b* Parameter Drift, Administration 1 Correlated Thetas of 0.6, Administration 2 Negatively Skewed (Shift) Distribution

		Aberrant Classification			
		1	2	3	4
True Classification	1	0.087	0.018	0.000	0.000
	2	0.027	0.207	0.103	0.000
	3	0.000	0.025	0.384	0.058
	4	0.000	0.000	0.005	0.087

		Delta Plot Purified Classification			
		1	2	3	4
True Classification	1	0.087	0.018	0.000	0.000
	2	0.028	0.208	0.102	0.000
	3	0.000	0.025	0.384	0.057
	4	0.000	0.000	0.005	0.087

		b Plot Purified Classification			
		1	2	3	4
True Classification	1	0.087	0.018	0.000	0.000
	2	0.027	0.208	0.102	0.000
	3	0.000	0.025	0.386	0.055
	4	0.000	0.000	0.005	0.086

		RPU Purified Classification			
		1	2	3	4
True Classification	1	0.093	0.012	0.000	0.000
	2	0.042	0.218	0.077	0.000
	3	0.000	0.037	0.379	0.050
	4	0.000	0.000	0.006	0.085

Continued, next page.

Table D.1., cont'd.:

Performance Level Contingency Tables, True vs. Aberrant & Purified Classification, 3 Spread Items with -0.7 *a* & -0.5 *b* Parameter Drift, Administration 1 Correlated Thetas of 0.6, Administration 2 Negatively Skewed (Shift) Distribution

		Aberrant Classification			
		1	2	3	4
True Classification	1	0.086	0.019	0.000	0.000
	2	0.025	0.203	0.110	0.000
	3	0.000	0.023	0.383	0.061
	4	0.000	0.000	0.004	0.087

		Delta Plot Purified Classification			
		1	2	3	4
True Classification	1	0.087	0.018	0.000	0.000
	2	0.028	0.208	0.102	0.000
	3	0.000	0.026	0.384	0.057
	4	0.000	0.000	0.005	0.087

		b Plot Purified Classification			
		1	2	3	4
True Classification	1	0.088	0.017	0.000	0.000
	2	0.029	0.209	0.099	0.000
	3	0.000	0.027	0.383	0.056
	4	0.000	0.000	0.005	0.086

		RPU Purified Classification			
		1	2	3	4
True Classification	1	0.093	0.012	0.000	0.000
	2	0.041	0.218	0.078	0.000
	3	0.000	0.036	0.380	0.050
	4	0.000	0.000	0.006	0.085

Continued, next page.

Table D.1., cont'd.:

Performance Level Contingency Tables, True vs. Aberrant & Purified Classification, 3 Spread Items with -0.7 *a* & -0.8 *b* Parameter Drift, Administration 1 Correlated Thetas of 0.6, Administration 2 Negatively Skewed (Shift) Distribution

		Aberrant Classification			
		1	2	3	4
True Classification	1	0.086	0.019	0.000	0.000
	2	0.025	0.203	0.109	0.000
	3	0.000	0.023	0.383	0.060
	4	0.000	0.000	0.004	0.087

		Delta Plot Purified Classification			
		1	2	3	4
True Classification	1	0.087	0.018	0.000	0.000
	2	0.028	0.208	0.102	0.000
	3	0.000	0.025	0.384	0.057
	4	0.000	0.000	0.005	0.087

		b Plot Purified Classification			
		1	2	3	4
True Classification	1	0.087	0.018	0.000	0.000
	2	0.027	0.207	0.103	0.000
	3	0.000	0.025	0.384	0.057
	4	0.000	0.000	0.005	0.087

		RPU Purified Classification			
		1	2	3	4
True Classification	1	0.093	0.012	0.000	0.000
	2	0.042	0.217	0.079	0.000
	3	0.000	0.036	0.379	0.052
	4	0.000	0.000	0.006	0.086

Continued, next page.

Table D.1., cont'd.:

Performance Level Contingency Tables, True vs. Aberrant & Purified Classification, 3 Moderate Items with -0.3 *a* & -0.5 *b* Parameter Drift, Administration 1 Correlated Thetas of 0.6, Administration 2 Negatively Skewed (Shift) Distribution

		Aberrant Classification			
		1	2	3	4
True Classification	1	0.087	0.018	0.000	0.000
	2	0.026	0.204	0.107	0.000
	3	0.000	0.024	0.381	0.061
	4	0.000	0.000	0.004	0.087

		Delta Plot Purified Classification			
		1	2	3	4
True Classification	1	0.087	0.018	0.000	0.000
	2	0.027	0.205	0.105	0.000
	3	0.000	0.024	0.382	0.060
	4	0.000	0.000	0.004	0.087

		b Plot Purified Classification			
		1	2	3	4
True Classification	1	0.087	0.018	0.000	0.000
	2	0.027	0.206	0.105	0.000
	3	0.000	0.025	0.383	0.058
	4	0.000	0.000	0.004	0.087

		RPU Purified Classification			
		1	2	3	4
True Classification	1	0.092	0.013	0.000	0.000
	2	0.039	0.218	0.081	0.000
	3	0.000	0.035	0.382	0.050
	4	0.000	0.000	0.006	0.085

Continued, next page.

Table D.1., cont'd.:

Performance Level Contingency Tables, True vs. Aberrant & Purified Classification, 3 Moderate Items with -0.7 *a* & -0.5 *b* Parameter Drift, Administration 1 Correlated Thetas of 0.6, Administration 2 Negatively Skewed (Shift) Distribution

		Aberrant Classification			
		1	2	3	4
True Classification	1	0.081	0.023	0.000	0.000
	2	0.019	0.197	0.122	0.000
	3	0.000	0.019	0.389	0.058
	4	0.000	0.000	0.005	0.087

		Delta Plot Purified Classification			
		1	2	3	4
True Classification	1	0.085	0.019	0.000	0.000
	2	0.025	0.207	0.105	0.000
	3	0.000	0.024	0.389	0.053
	4	0.000	0.000	0.005	0.086

		b Plot Purified Classification			
		1	2	3	4
True Classification	1	0.086	0.019	0.000	0.000
	2	0.026	0.208	0.104	0.000
	3	0.000	0.025	0.387	0.055
	4	0.000	0.000	0.005	0.086

		RPU Purified Classification			
		1	2	3	4
True Classification	1	0.091	0.014	0.000	0.000
	2	0.036	0.220	0.081	0.000
	3	0.000	0.035	0.388	0.044
	4	0.000	0.000	0.007	0.084

Continued, next page.

Table D.1., cont'd.:

Performance Level Contingency Tables, True vs. Aberrant & Purified Classification, 3 Moderate Items with -0.7 *a* & -0.5 *b* Parameter Drift, Administration 1 Correlated Thetas of 0.6, Administration 2 Negatively Skewed (Shift) Distribution

		Aberrant Classification			
		1	2	3	4
True Classification	1	0.085	0.019	0.000	0.000
	2	0.024	0.200	0.113	0.000
	3	0.000	0.022	0.380	0.064
	4	0.000	0.000	0.004	0.088

		Delta Plot Purified Classification			
		1	2	3	4
True Classification	1	0.087	0.018	0.000	0.000
	2	0.027	0.205	0.105	0.000
	3	0.000	0.025	0.381	0.061
	4	0.000	0.000	0.004	0.087

		b Plot Purified Classification			
		1	2	3	4
True Classification	1	0.087	0.017	0.000	0.000
	2	0.028	0.207	0.102	0.000
	3	0.000	0.025	0.382	0.059
	4	0.000	0.000	0.004	0.087

		RPU Purified Classification			
		1	2	3	4
True Classification	1	0.092	0.013	0.000	0.000
	2	0.040	0.217	0.080	0.000
	3	0.000	0.035	0.381	0.050
	4	0.000	0.000	0.006	0.085

Continued, next page.

Table D.1., cont’d.:

Performance Level Contingency Tables, True vs. Aberrant & Purified Classification, 3 Moderate Items with -0.7 *a* & -0.8 *b* Parameter Drift, Administration 1 Correlated Thetas of 0.6, Administration 2 Negatively Skewed (Shift) Distribution

		Aberrant Classification			
		1	2	3	4
True Classification	1	0.081	0.024	0.000	0.000
	2	0.018	0.196	0.123	0.000
	3	0.000	0.019	0.389	0.059
	4	0.000	0.000	0.004	0.087

		Delta Plot Purified Classification			
		1	2	3	4
True Classification	1	0.085	0.020	0.000	0.000
	2	0.024	0.206	0.107	0.000
	3	0.000	0.024	0.388	0.054
	4	0.000	0.000	0.005	0.086

		b Plot Purified Classification			
		1	2	3	4
True Classification	1	0.086	0.019	0.000	0.000
	2	0.025	0.207	0.105	0.000
	3	0.000	0.024	0.387	0.055
	4	0.000	0.000	0.005	0.086

		RPU Purified Classification			
		1	2	3	4
True Classification	1	0.093	0.012	0.000	0.000
	2	0.041	0.219	0.077	0.000
	3	0.000	0.037	0.382	0.048
	4	0.000	0.000	0.006	0.085

Continued, next page.

Table D.1., cont'd.:

Performance Level Contingency Tables, True vs. Aberrant & Purified Classification, 5 Spread Items with -0.3 *a* & -0.5 *b* Parameter Drift, Administration 1 Correlated Thetas of 0.6, Administration 2 Negatively Skewed (Shift) Distribution

		Aberrant Classification			
		1	2	3	4
True Classification	1	0.084	0.020	0.000	0.000
	2	0.023	0.203	0.112	0.000
	3	0.000	0.022	0.386	0.058
	4	0.000	0.000	0.005	0.087

		Delta Plot Purified Classification			
		1	2	3	4
True Classification	1	0.085	0.019	0.000	0.000
	2	0.025	0.205	0.108	0.000
	3	0.000	0.024	0.386	0.057
	4	0.000	0.000	0.005	0.087

		b Plot Purified Classification			
		1	2	3	4
True Classification	1	0.085	0.020	0.000	0.000
	2	0.024	0.205	0.109	0.000
	3	0.000	0.023	0.387	0.056
	4	0.000	0.000	0.005	0.086

		RPU Purified Classification			
		1	2	3	4
True Classification	1	0.092	0.013	0.000	0.000
	2	0.039	0.217	0.081	0.000
	3	0.000	0.035	0.382	0.050
	4	0.000	0.000	0.006	0.085

Continued, next page.

Table D.1., cont'd.:

Performance Level Contingency Tables, True vs. Aberrant & Purified Classification, 5 Spread Items with -0.7 *a* & -0.5 *b* Parameter Drift, Administration 1 Correlated Thetas of 0.6, Administration 2 Negatively Skewed (Shift) Distribution

		Aberrant Classification			
		1	2	3	4
True Classification	1	0.079	0.026	0.000	0.000
	2	0.017	0.202	0.118	0.000
	3	0.000	0.020	0.402	0.044
	4	0.000	0.000	0.007	0.084

		Delta Plot Purified Classification			
		1	2	3	4
True Classification	1	0.085	0.020	0.000	0.000
	2	0.024	0.211	0.102	0.000
	3	0.000	0.026	0.394	0.046
	4	0.000	0.000	0.007	0.085

		b Plot Purified Classification			
		1	2	3	4
True Classification	1	0.085	0.020	0.000	0.000
	2	0.024	0.211	0.103	0.000
	3	0.000	0.026	0.394	0.046
	4	0.000	0.000	0.007	0.085

		RPU Purified Classification			
		1	2	3	4
True Classification	1	0.089	0.016	0.000	0.000
	2	0.032	0.225	0.081	0.000
	3	0.000	0.036	0.397	0.034
	4	0.000	0.000	0.010	0.081

Continued, next page.

Table D.1., cont'd.:

Performance Level Contingency Tables, True vs. Aberrant & Purified Classification, 5 Spread Items with -0.7 *a* & -0.5 *b* Parameter Drift, Administration 1 Correlated Thetas of 0.6, Administration 2 Negatively Skewed (Shift) Distribution

		Aberrant Classification			
		1	2	3	4
True Classification	1	0.082	0.023	0.000	0.000
	2	0.020	0.196	0.122	0.000
	3	0.000	0.019	0.385	0.062
	4	0.000	0.000	0.004	0.087

		Delta Plot Purified Classification			
		1	2	3	4
True Classification	1	0.085	0.020	0.000	0.000
	2	0.024	0.202	0.111	0.000
	3	0.000	0.022	0.384	0.060
	4	0.000	0.000	0.004	0.087

		b Plot Purified Classification			
		1	2	3	4
True Classification	1	0.086	0.019	0.000	0.000
	2	0.025	0.205	0.107	0.000
	3	0.000	0.024	0.385	0.058
	4	0.000	0.000	0.005	0.087

		RPU Purified Classification			
		1	2	3	4
True Classification	1	0.092	0.013	0.000	0.000
	2	0.040	0.217	0.080	0.000
	3	0.000	0.035	0.381	0.050
	4	0.000	0.000	0.006	0.085

Continued, next page.

Table D.1., cont'd.:

Performance Level Contingency Tables, True vs. Aberrant & Purified Classification, 5 Spread Items with -0.7 *a* & -0.8 *b* Parameter Drift, Administration 1 Correlated Thetas of 0.6, Administration 2 Negatively Skewed (Shift) Distribution

		Aberrant Classification			
		1	2	3	4
True Classification	1	0.077	0.027	0.000	0.000
	2	0.015	0.198	0.125	0.000
	3	0.000	0.019	0.402	0.046
	4	0.000	0.000	0.007	0.084

		Delta Plot Purified Classification			
		1	2	3	4
True Classification	1	0.083	0.022	0.000	0.000
	2	0.022	0.207	0.109	0.000
	3	0.000	0.024	0.394	0.049
	4	0.000	0.000	0.006	0.085

		b Plot Purified Classification			
		1	2	3	4
True Classification	1	0.084	0.021	0.000	0.000
	2	0.022	0.207	0.108	0.000
	3	0.000	0.024	0.394	0.049
	4	0.000	0.000	0.006	0.085

		RPU Purified Classification			
		1	2	3	4
True Classification	1	0.091	0.014	0.000	0.000
	2	0.038	0.225	0.074	0.000
	3	0.000	0.039	0.391	0.036
	4	0.000	0.000	0.009	0.082

APPENDIX E

CLASSIFICATION PLOTS

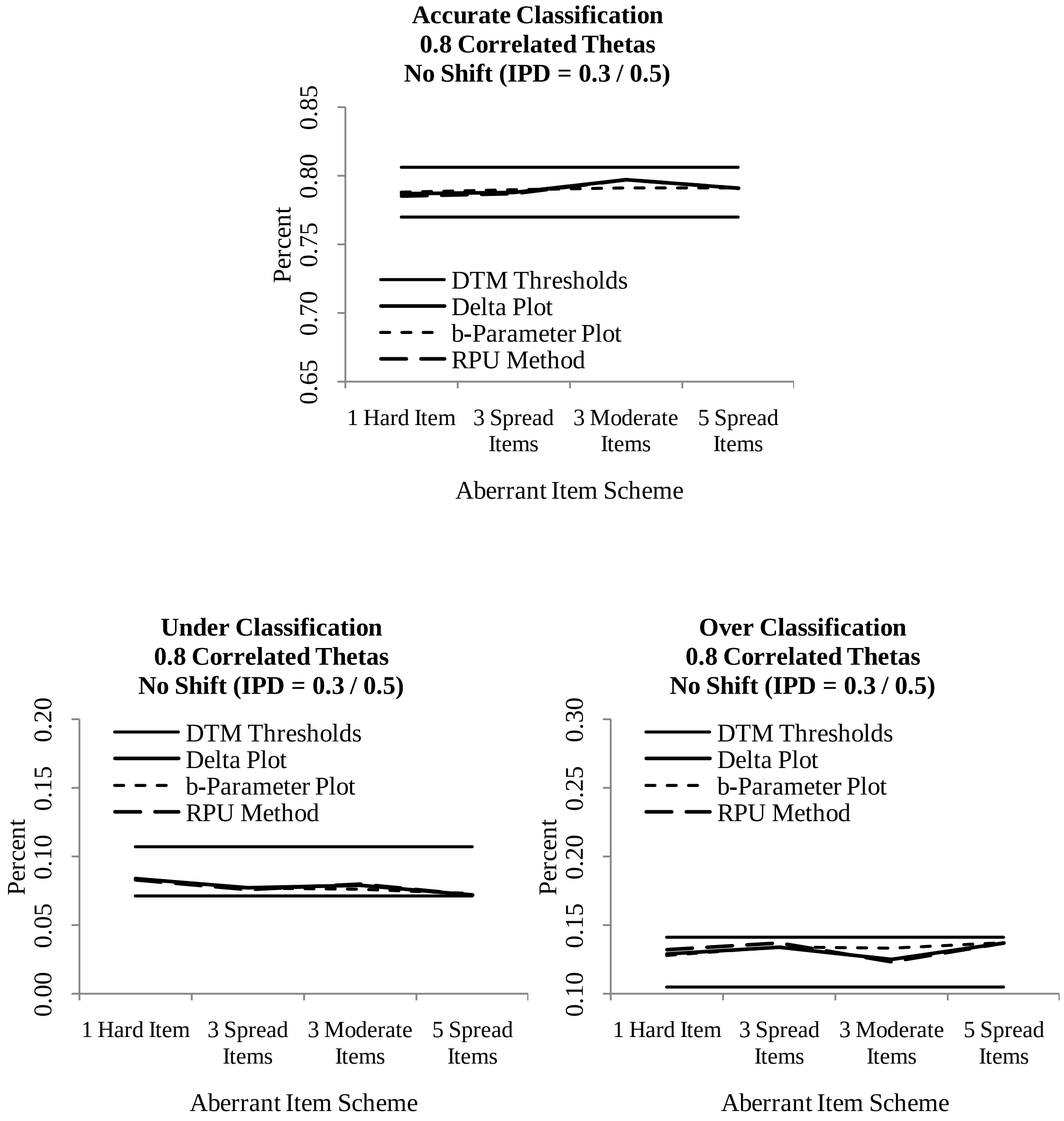


Figure E.1. Classification Plots

Continued, next page.

Figure E.1., cont'd.:

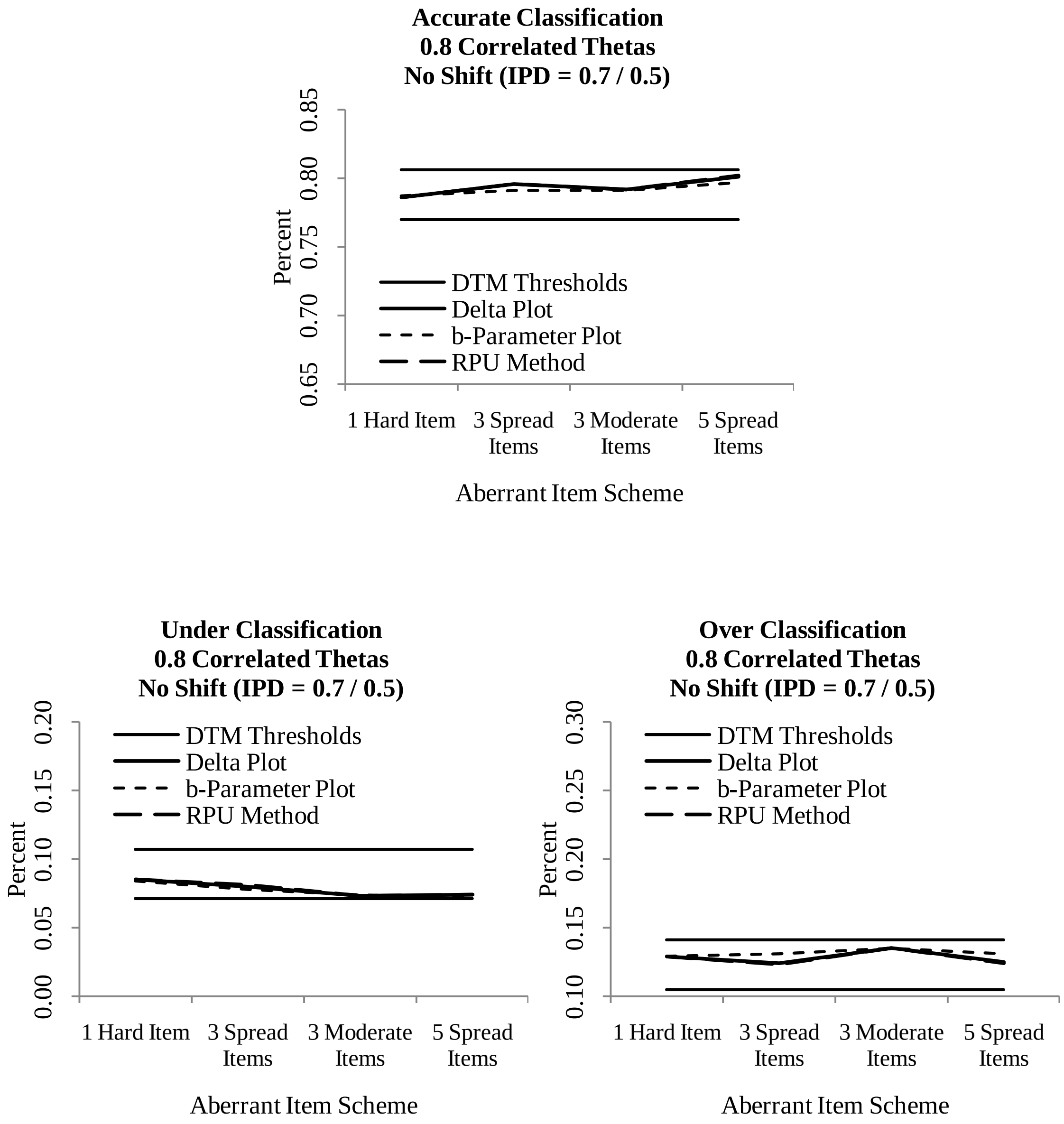
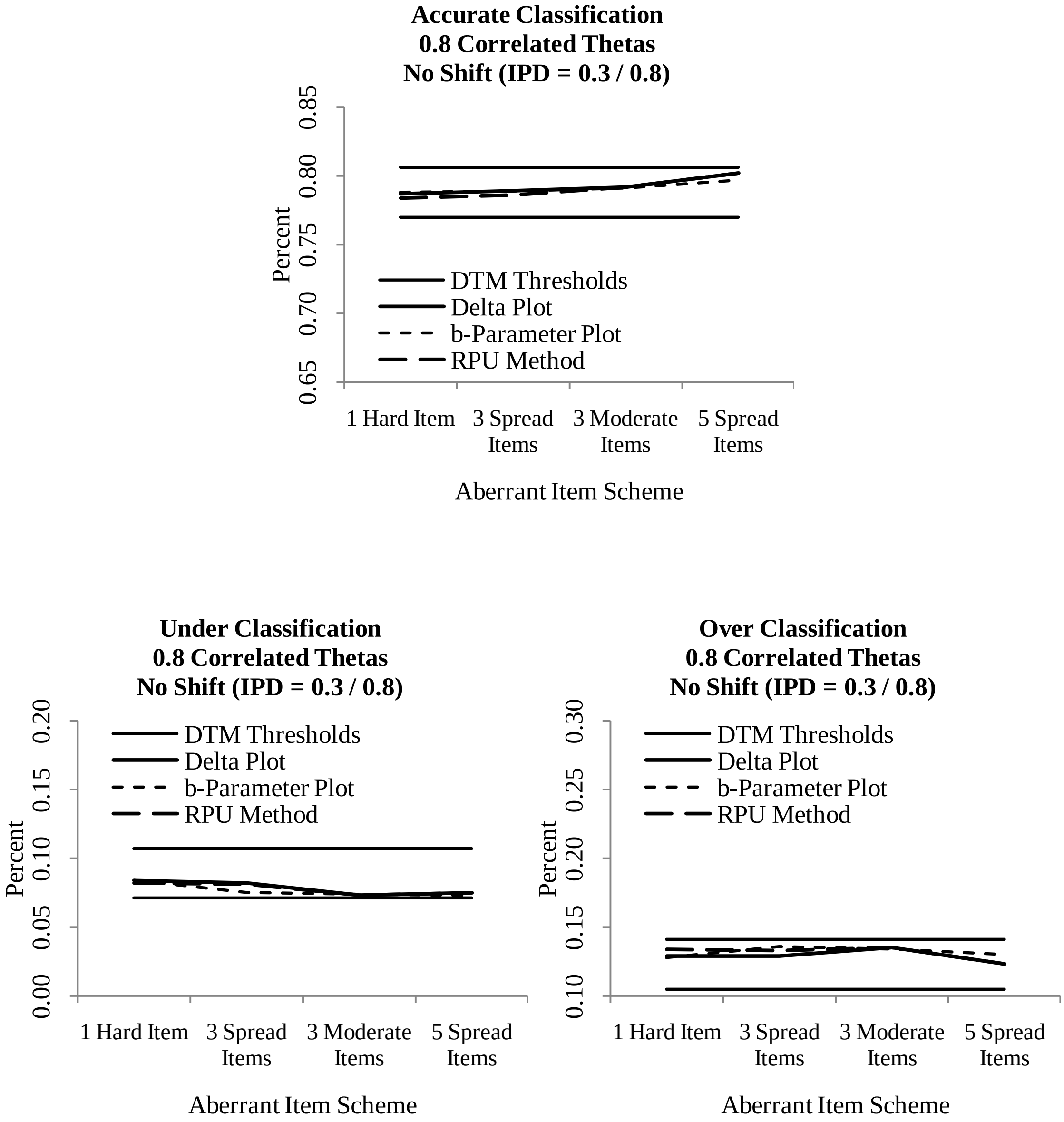


Figure E.1. Classification Plots

Continued, next page.

Figure E.1., cont’d.:


Accurate Classification 0.8 Correlated Thetas No Shift (IPD = 0.3 / 0.8)

Under Classification 0.8 Correlated Thetas No Shift (IPD = 0.3 / 0.8)

Over Classification 0.8 Correlated Thetas No Shift (IPD = 0.3 / 0.8)

Figure E.1. Classification Plots

Continued, next page.

Figure E.1., cont'd.:

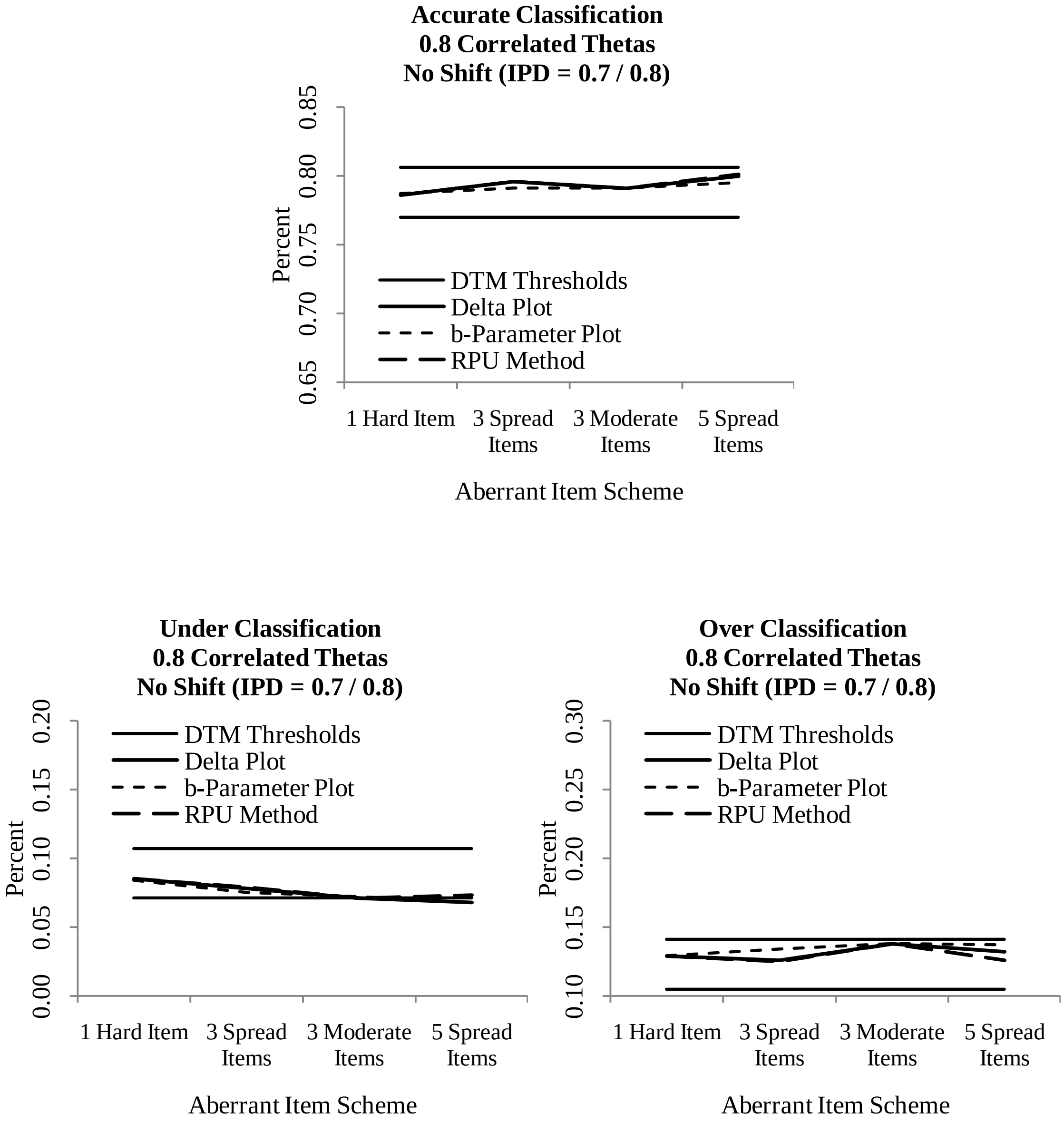
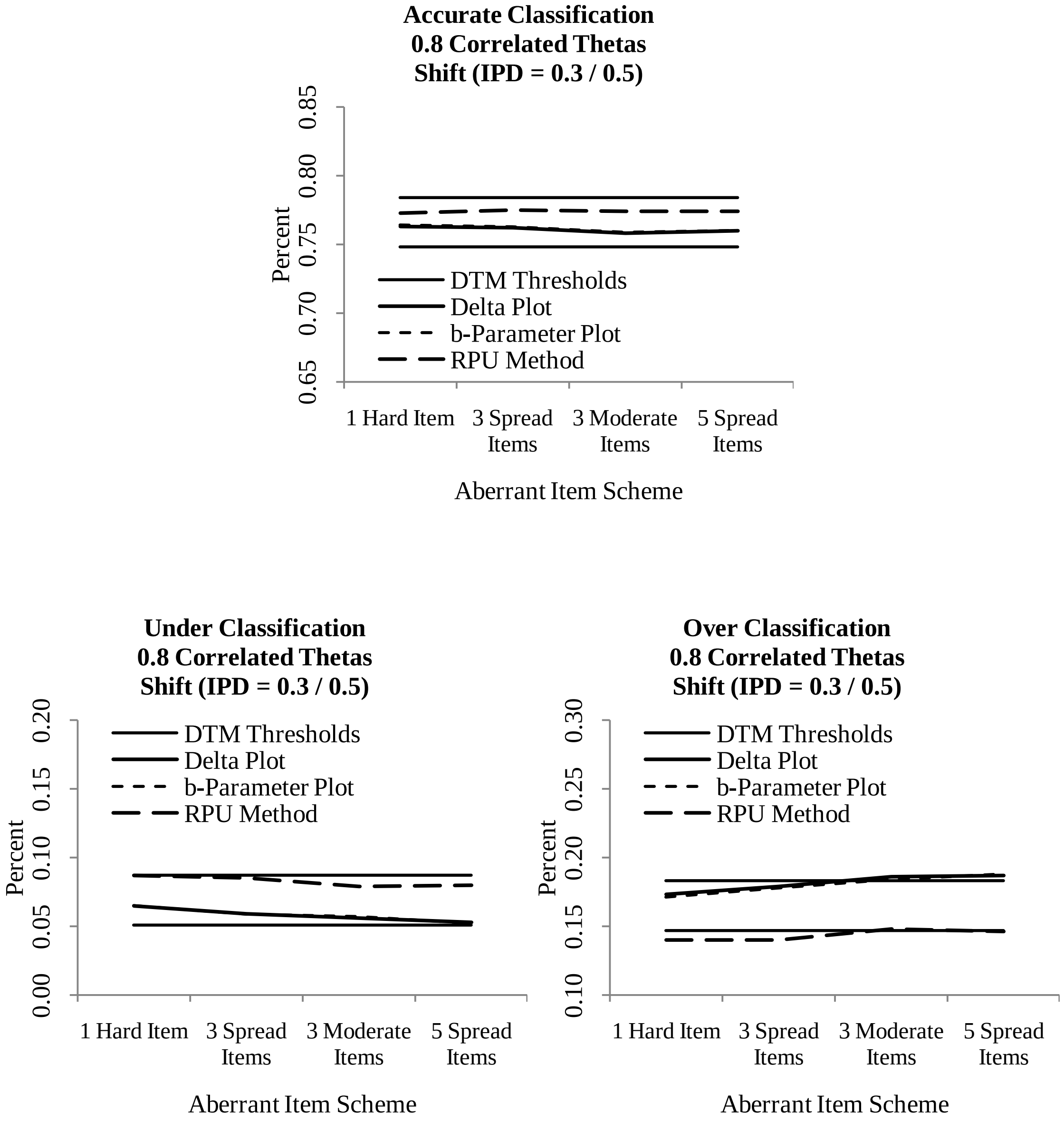


Figure E.1. Classification Plots

Continued, next page.

Figure E.1., cont'd.:


Accurate Classification 0.8 Correlated Thetas Shift (IPD = 0.3 / 0.5)

Under Classification 0.8 Correlated Thetas Shift (IPD = 0.3 / 0.5)

Over Classification 0.8 Correlated Thetas Shift (IPD = 0.3 / 0.5)

Figure E.1. Classification Plots

Continued, next page.

Figure E.1., cont'd.:

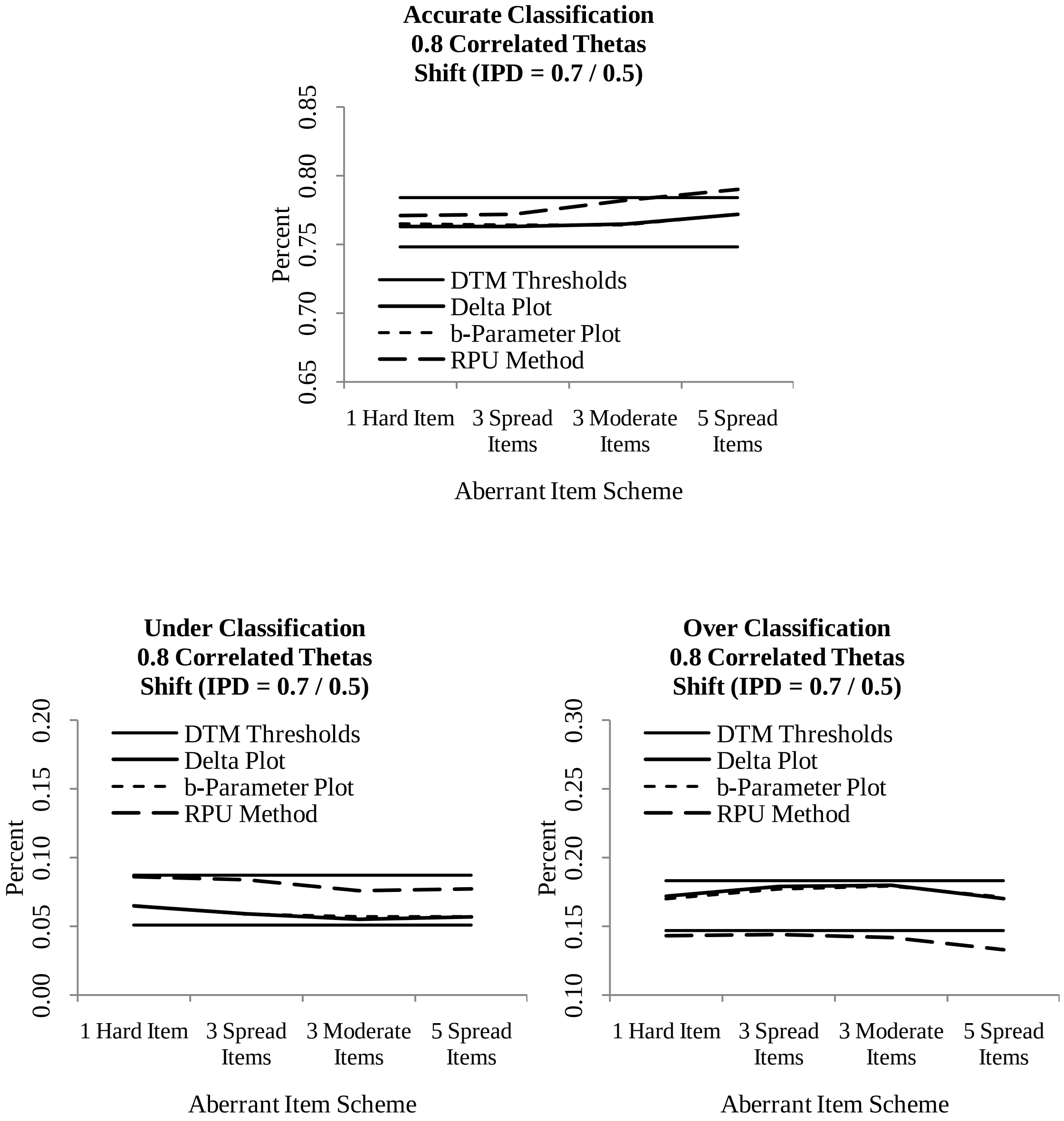
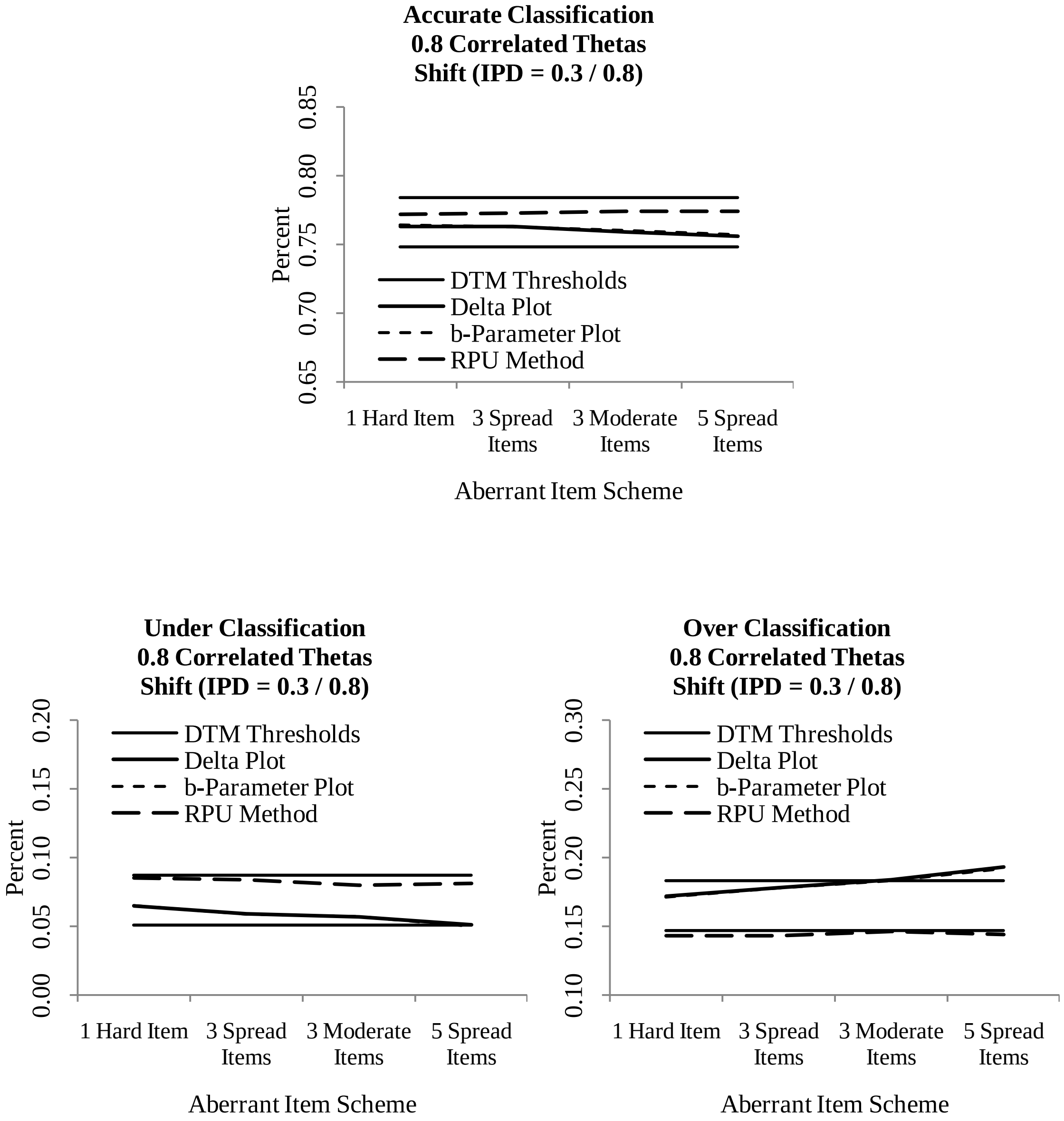


Figure E.1. Classification Plots

Continued, next page.

Figure E.1., cont'd.:



Accurate Classification
0.8 Correlated Thetas
Shift (IPD = 0.3 / 0.8)

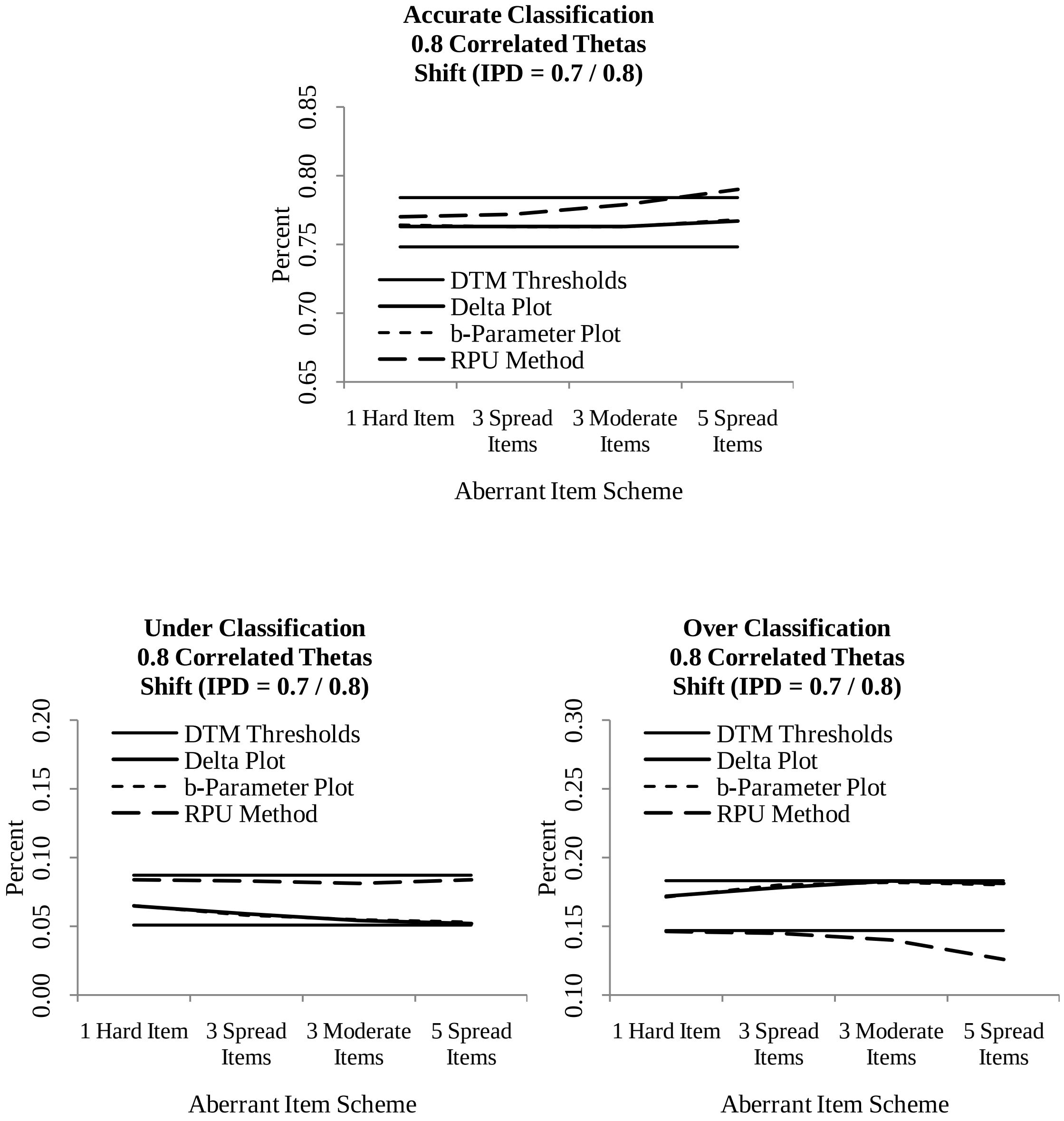
Under Classification
0.8 Correlated Thetas
Shift (IPD = 0.3 / 0.8)

Over Classification
0.8 Correlated Thetas
Shift (IPD = 0.3 / 0.8)

Figure E.1. Classification Plots

Continued, next page.

Figure E.1., cont'd.:



Accurate Classification
0.8 Correlated Thetas
Shift (IPD = 0.7 / 0.8)

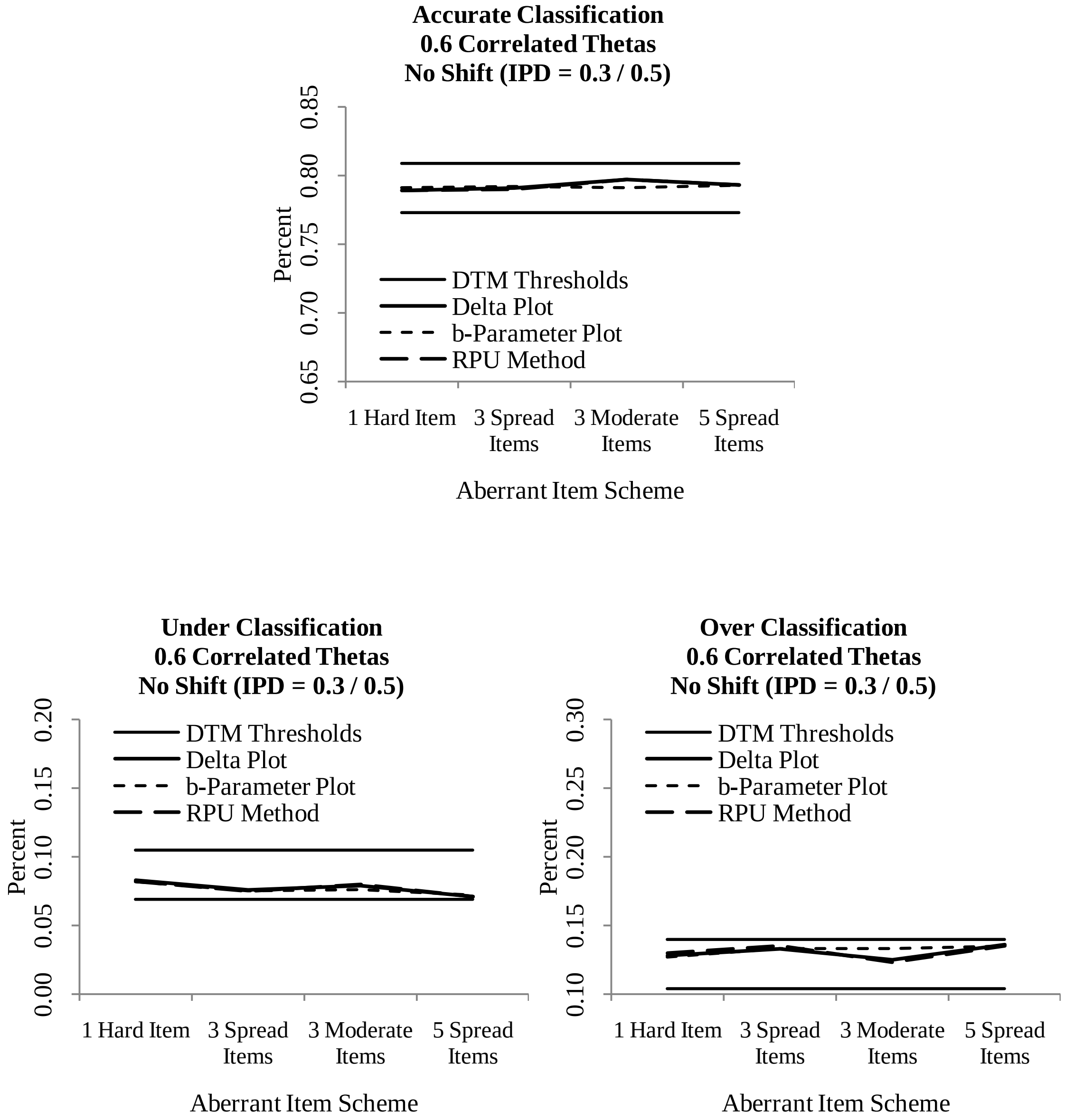
Under Classification
0.8 Correlated Thetas
Shift (IPD = 0.7 / 0.8)

Over Classification
0.8 Correlated Thetas
Shift (IPD = 0.7 / 0.8)

Figure E.1. Classification Plots

Continued, next page.

Figure E.1., cont'd.:


Accurate Classification
0.6 Correlated Thetas
No Shift (IPD = 0.3 / 0.5)

Under Classification
0.6 Correlated Thetas
No Shift (IPD = 0.3 / 0.5)

Over Classification
0.6 Correlated Thetas
No Shift (IPD = 0.3 / 0.5)

Figure E.1. Classification Plots

Continued, next page.

Figure E.1., cont'd.:

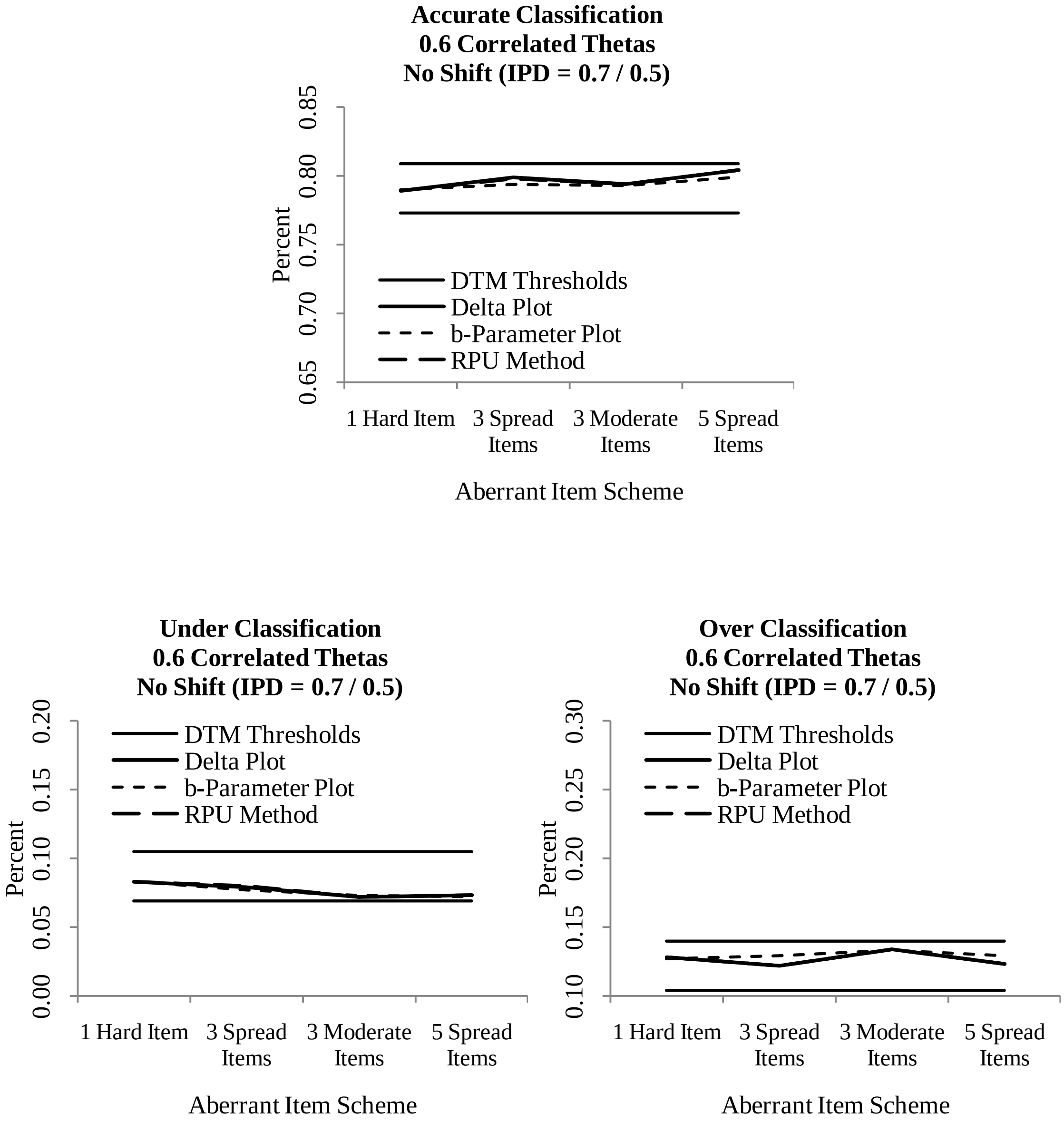


Figure E.1. Classification Plots

Continued, next page.

Figure E.1., cont'd.:

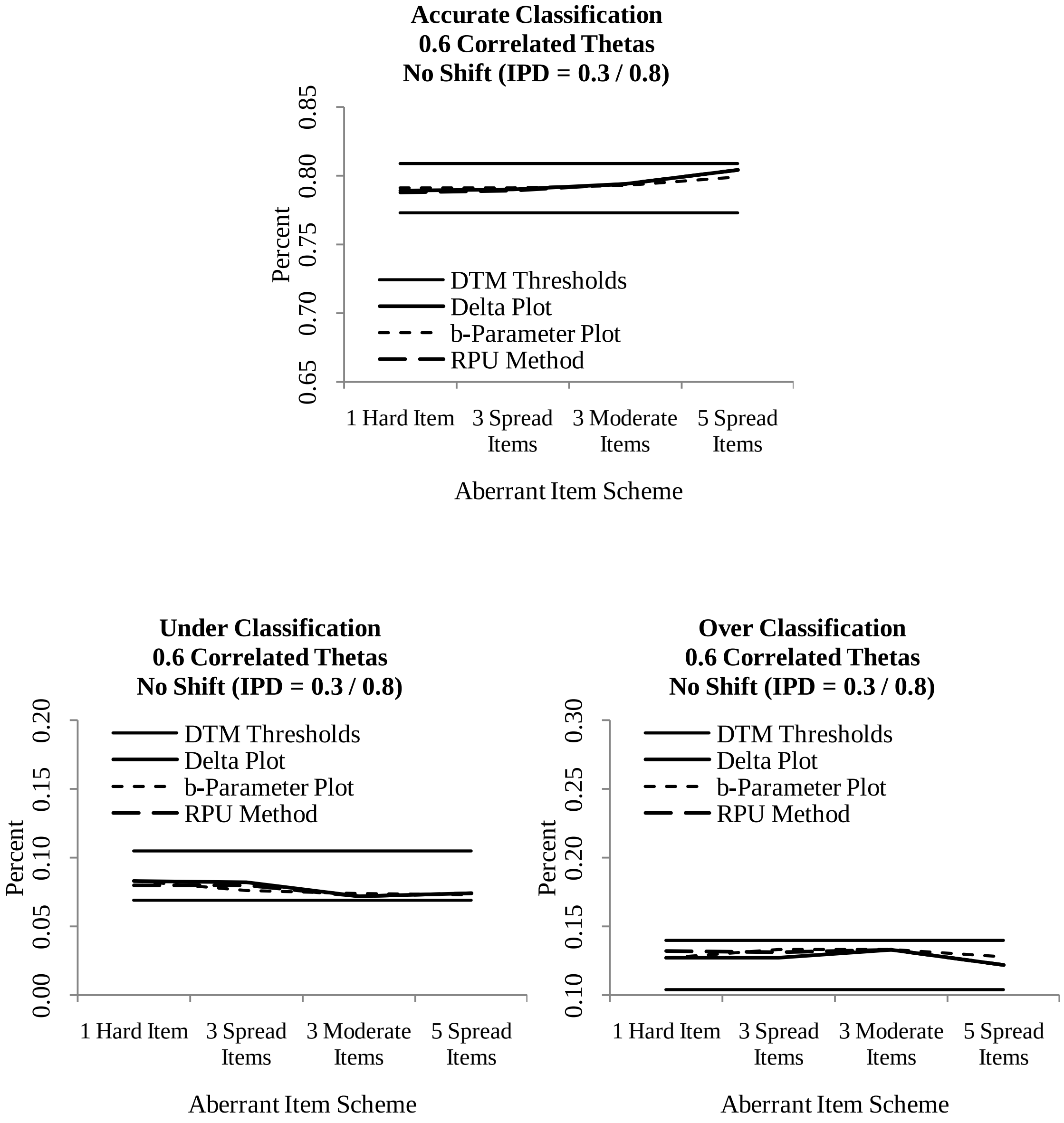


Figure E.1. Classification Plots

Continued, next page.

Figure E.1., cont'd.:

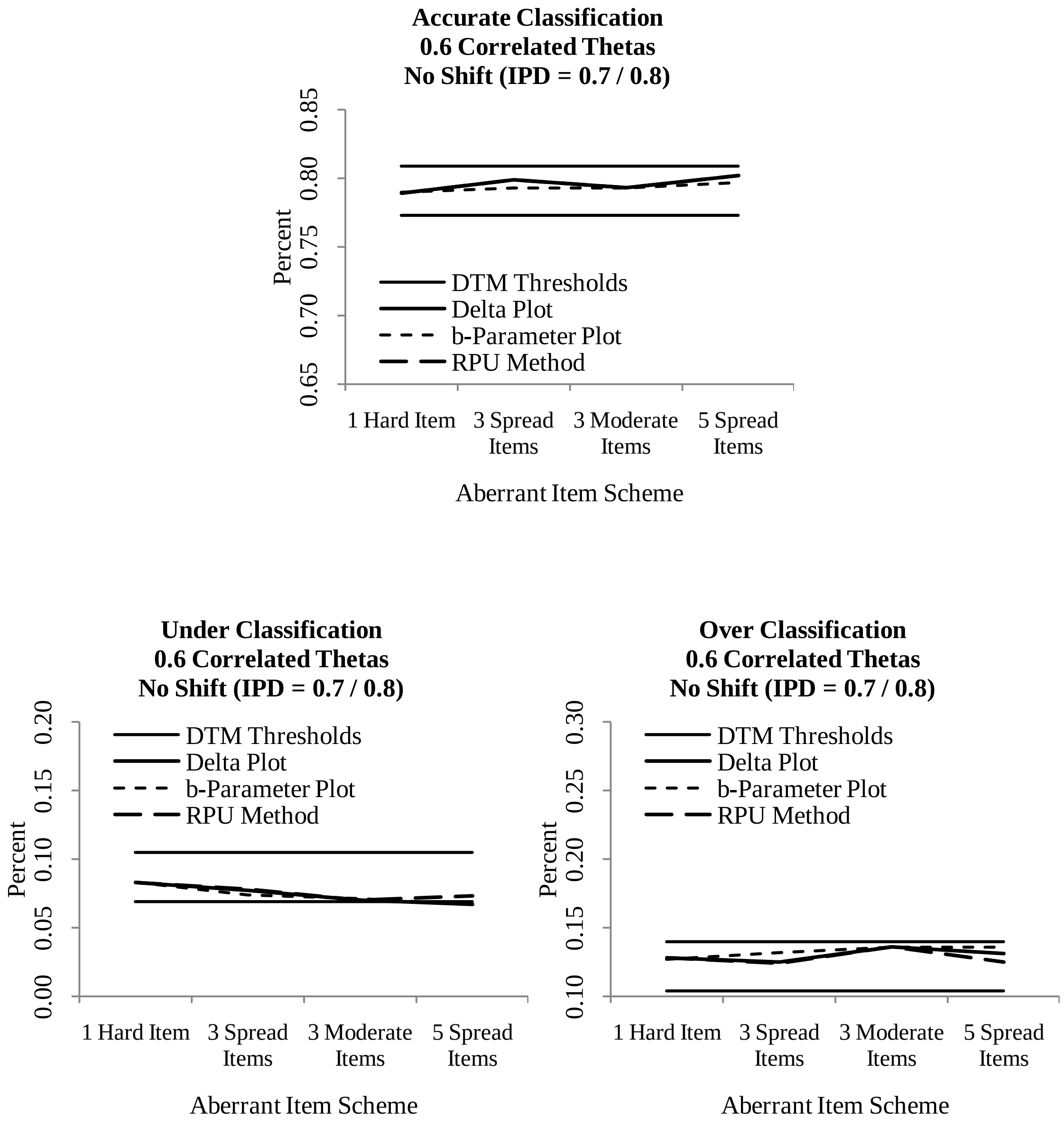


Figure E.1. Classification Plots

Continued, next page.

Figure E.1., cont'd.:

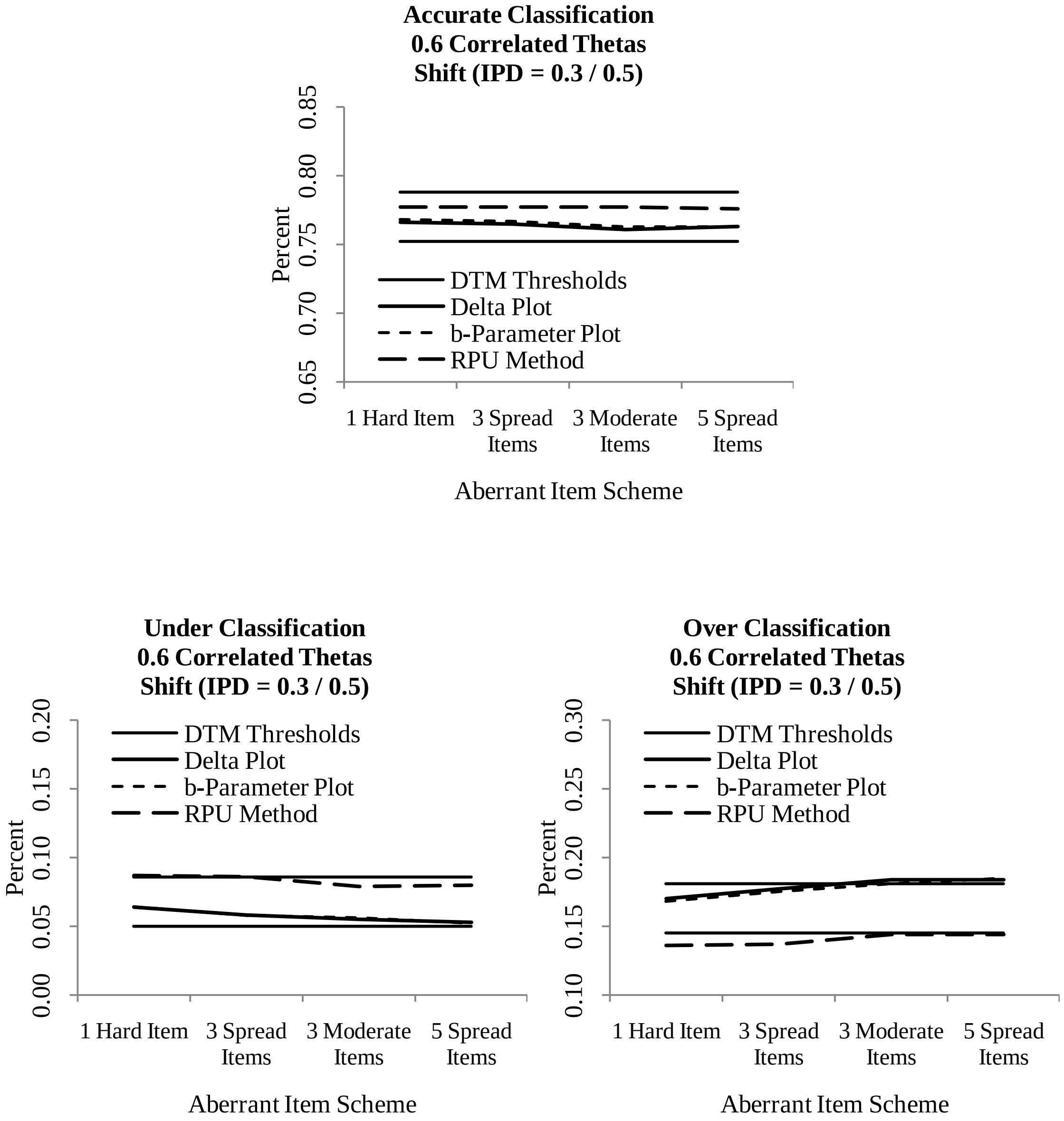


Figure E.1. Classification Plots

Continued, next page.

Figure E.1., cont'd.:

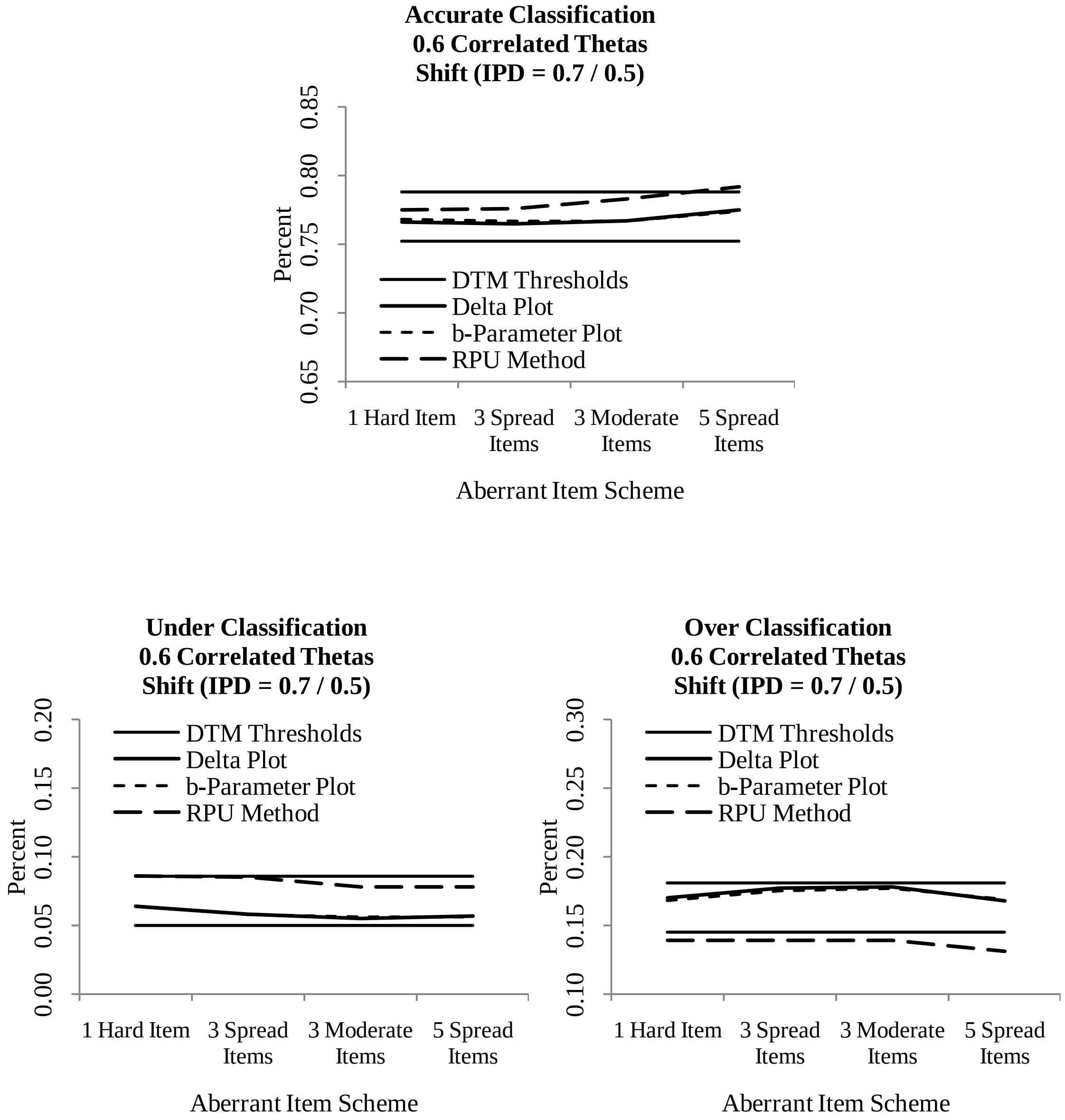
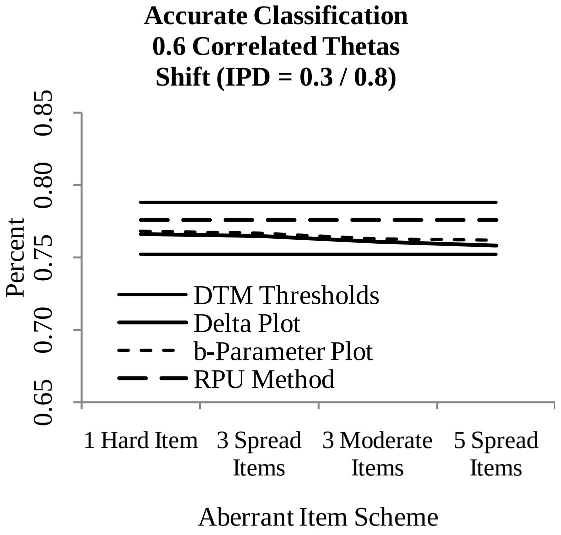


Figure E.1. Classification Plots

Continued, next page.

Figure E.1., cont'd.:



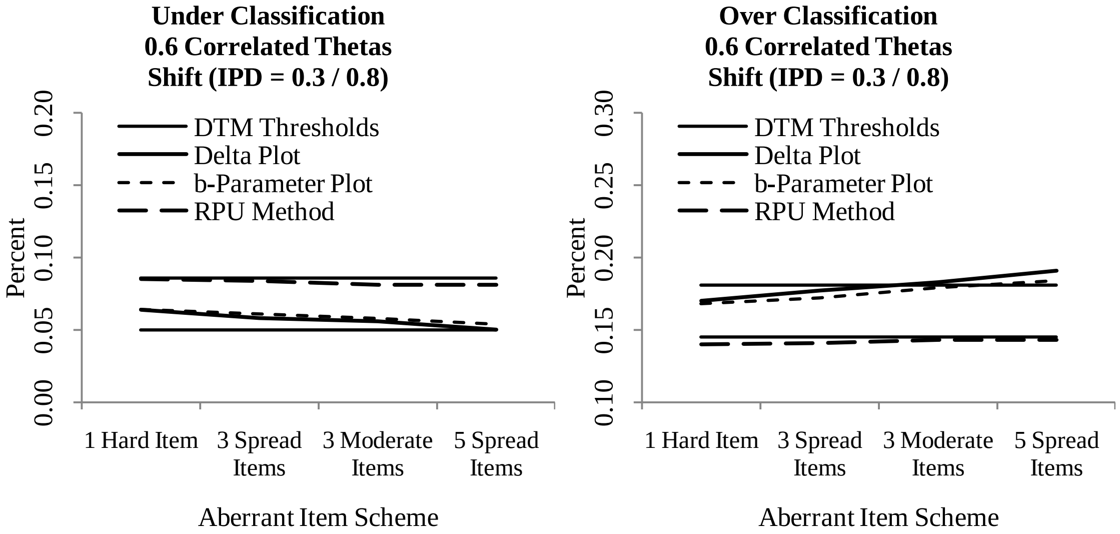


Figure E.1. Classification Plots

Continued, next page.

Figure E.1., cont'd.:

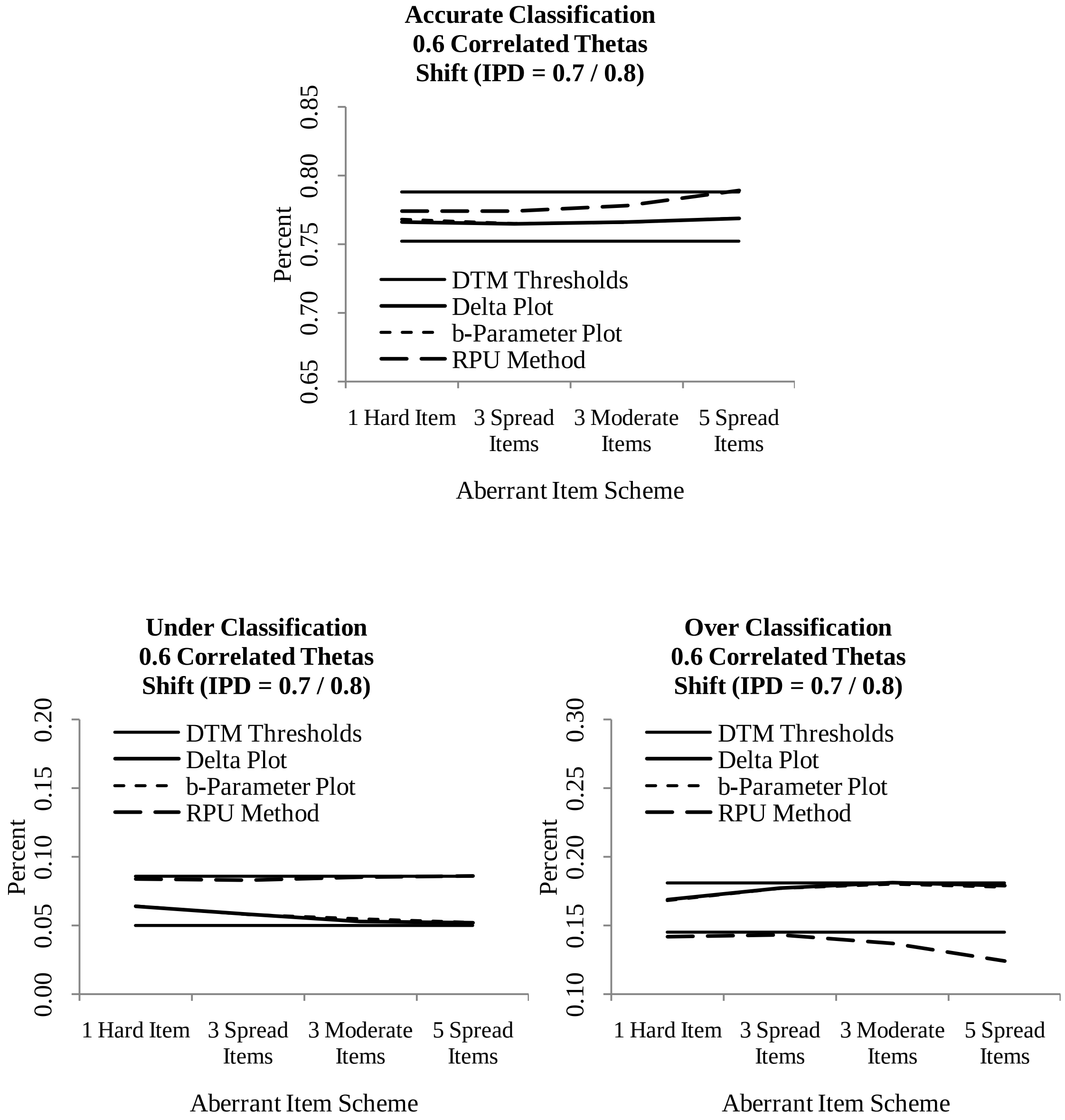


Figure E.1. Classification Plots

BIBLIOGRAPHY

Angoff, W. H. (1968). How we calibrate College Board scores. *College Board Review, 68*, 11-14.

Angoff, W. H. (1971). Scales, norms and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement, 4th* ed. (pp. 508-600). Washington, DC: American Council on Education.

Angoff, W. H., & Ford, S. F. (1973). Item-race interaction on a test of scholastic aptitude. *Journal of Educational Measurement, 10*, 107-116.

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement, 29*(1), 67-91.

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

Ankenmann, R. D., Witt, E. A., & Dunbar, S. B. (1999). An investigation of the power of the likelihood ratio goodness-of-fit statistic in detecting differential item functioning. *Journal of Educational Measurement, 36*, 277-300.

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

Brennan, R. (Ed.). (2006). Perspectives on the evolution and future of educational measurement. In R. Brennan (Ed.). *Educational measurement*, 4th ed. (pp. 1-16). Westport, CT: Praeger Publishers.

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

Camilli, G. (1979). *A critique of the chi-square method for assessing item bias*. Boulder, CO: Laboratory of Educational Research, University of Colorado.

Camilli, G. (2006). Test fairness. Brennan, R. L. (Ed.), *Educational measurement,* 4th ed. (pp. 221-256).

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items.* Thousand Oaks, CA: Sage.

Cook, L. L., & Eignor, D. R. (1989). Using item response theory in test score equating. *International Journal of Educational Research, 13*(2), 161-173.

Cook, L. L., & Eignor, D. R. (1991). IRT equating methods. *Educational Measurement: Issues and Practice, 10,* 37-45.

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.

Cronbach, L. J., & Warrington, W. G. (1952). Efficacy of multiple choice tests as a function of spread of item difficulties. *Psychometrika, 17*(2), 127-147.

de Ayala, R. J. (1999). *The theory and practice of item response theory*. New York: The Guilford Press.

Donoghue, J. R., & Isham, S. P. (1998). A comparison of procedures to detect item parameter drift. *Applied Psychological Measurement, 22(1),* 33-51.

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, Inc.

Dorans, N. J., Kubiak, A., & Melican, G. J. (1998). *Guidelines for selection of embedded common items for score equating* (ETS SR-98-02). Princeton, NJ: Educational Testing Service.

Eignor, D. R. (1985). *An investigation of the feasibility and practical outcomes of pre-equating the SAT verbal and mathematical sections* (RR-85-10). Princeton, NJ: Educational Testing Service.

Fitzpatrick, A. R. (2008). NCME 2008 presidential address: The impact of anchor set configuration on student proficiency rates. *Educational Measurement: Issues and Practice, 27*(4), 34-40.

Gierl, M. J., Jodoin, M. G., & Ackerman, T. A. (2000, March). *Performance of Mantel-Haenszel, simultaneous item bias test, and logistic regression when the proportion of DIF items is large.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, Louisiana.

Gulliksen, H. (1950). *Theory of mental tests*. New York: John Wiley.

Hambleton, R. K., & Rogers, H. J. (1989). Detecting potentially biased test items: Comparison of the IRT area and Mantel-Haenszel methods. *Applied Measurement in Education, 2,* 313-334.

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

Han, K. & Wells, C. S. (2007, April). Impact of differential item functioning (DIF) on test equating and proficiency estimates. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Hanson, B. A. & Béguin, A. A. (2002). Obtaining a common scale for Item Response Theory item parameters using separate versus concurrent estimation in the common-item equating design. *Applied Psychological Measurement*, *26,* 3-24.

Hills, J. R. (1989). Screening for potential biased items in testing programs. *Educational Measurement: Issues and Practice, 8,* 5-11.

Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. Brennan (Ed.), *Educational measurement,* 4th ed. (pp. 187-220). Westport, CT: Praeger Publishers.

Holland, P. W., & Thayer, D. T. (1985). *An alternative definition of the ETS delta scale of item difficulty* (Research Rep. No. 85-43). Princeton, NJ: Educational Testing Service.

Holland, P. W., & Thayer, D. T. (1998). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum.

Hu, H., Rogers, W. T., & Vukmirovic, Z. (2008). Investigation of IRT-based equating methods in the presence of outlier common items. *Applied Psychological Measurement, 32*(4), 311-333.

Huff, K. L., & Hambleton, R. K. (2001). *The detection and exclusion of differentially functioning anchor items* (Research Report 415). Amherst, MA: Laboratory of Psychometric and Evaluation, University of Massachusetts.

Ironson, G.H., & Subkoviak, M. (1979). A comparison of several methods of assessing item bias. *Journal of Educational Measurement,16,* 209-225.

Jodoin, M. G. & Gierl, M. J. (2001). Evaluating Type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education, 14*(4), 329-349.

Jodoin, M. G., Keller, L. A., & Swaminathan, H. (2003). A comparison of linear, fixed common item, and concurrent parameter estimation equating procedures in capturing academic growth. *Journal of Experimental Education*, *71*, 229-250.

Karkee, T., & Choi, S. (2005, April). *Impact of eliminating anchor items flagged from statistical criteria on test score classifications in common item equating.* Paper presented at the meeting of the American Educational Research Association, Montreal, Canada.

Keller, L. A., Egan, K. L., & Schneider, M. C. (2010). *Item parameter drift in anchor items-detection and consequences: An analysis of simulated and operational test data.* CTB/McGraw-Hill: Monterey, CA.

Keller, L., Wells, C. (2009). *The effect of differentially functioning anchor items on the classification of examinees*. Paper presented at the meeting of the National Council on Measurement in Education, San Diego, CA

Kim, S., & Cohen, A. S. (1995). A comparison of Lord's chi-square, Raju's area measures, and the likelihood ratio test on detection of differential item functioning. *Applied Measurement in Education, 8*(4), 291-312.

Kim, S.-H., & Cohen, A. S. (1997, March). *An investigation of the likelihood ratio test for detection of differential item functioning under the graded response model.* Paper presented at the meeting of the American Educational Research Association, Chicago, IL.

Kim, S., & Kolen, M. J. (2004). *STUIRT* [Computer software]. Iowa City, IA: Iowa Testing Programs, The University of Iowa.

Kim, W. & Nering, M. (2007). *Evaluation of equating items using DFIT.* Paper presented at the meeting of the National Council on Measurement in Education, Chicago, IL.

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.

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

Kolen, M. J. (1988). Traditional equating methodology. *Educational Measurement: Issues and Practice, 7*(4), 29-37.

Kolen, M. J. (2004). Population invariance in equating and linking: Concept and history. *Journal of Educational Measurement, 41*(1), 3-14.

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

Kristjansson, E., Aylesworth, R., McDowell, I., & Zumbo, B. D. (2005). A comparison of four methods for detecting differential item functioning in ordered response items. *Educational and Psychological Measurement, 65*(6), 935-953.

Linn, R. L., & Harnisch, D. L. (1981). Interactions between item content and group membership. *Journal of Educational Measurement, 18,* 109-118.

Longford, N. T., Holland, P. W., & Thayer, D. T. (1993). Stability of the MH DIF across populations. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 171-196). Hillside, NJ: Erlbaum.

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

Marco, G. L., Petersen, N. S., & Stewart, E. E. (1983). A test of the adequacy of curvilinear score equating models. In D. Weiss (Ed.), *New horizons in testing: Latent trait theory and computerized adaptive testing* (pp. 147-177). New York: Academic Press.

Mazor, K. M., Clauser, B. E. & Hambleton, R. K. (1994). Identification of nonuniform differential item functioning using a variation of the Mantel_Haenszel Procedure. *Educational and Psychological Measurement,* 54, 284-291.

McKinley, R. L., & Reckase, M. D. (1981). *A comparison of procedures for constructing large item pools* (Research Report 81-3). Columbia, MO: Department of Educational Psychology, University of Missouri.

Merz, W.R., & Grossen, N.E. (1979). *An empirical investigation of six methods for examining test item bias* (Research Report NIE-6-78-0067). Sacramento, CA: National Institute of Education, California State University.

Michaelides, M. (2006). *Effects of misbehaving common items on aggregate scores and an application of the Mantel-Haenszel statistic in test equating* (CSE Report 688). Los Angeles, CA: Center for the Study of Evaluation, University of California.

Michaelides, M. P. (2008). An illustration of a Mantel-Haenszel procedure to flag misbehaving common items in test equating. *Practical Assessment Research* & *Evaluation, 13*(7). Available online: http://pareonline.net/getvn.asp?v=13&n=7

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

Narayanan, P. & Swaminathan, H. (1996). Identification of Items that Show Nonuniform DIF *Applied Psychological Measurement*, 20(3), 257-274.

No Child Left Behind Act of 2001, Public Law No. 107-110, 115 Stat. 1425.

Oshima, T. C., & Morris, S. B. (2008). An NCME instructional module on Raju's differential functioning of items and tests (DFIT). *Educational Measurement: Issues and Practice, 27,* 43-50.

PARSCALE [Computer software]. (2003). Lincolnwood, IL: Scientific Software International, Inc.

Penfield, R. D. (2001). Assessing differential item functioning across multiple groups: A comparison of three Mantel-Haenszel procedures. *Applied Measurement in Education, 14*, 235-259.

Penfield, R. D., & Lam, T. C. (2000). Assessing differential item functioning in performance assessment: Review and recommendations. *Educational Measurement: Issues and Practice, 19*, 5-15.

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. In R. L. Linn (Ed.), *Educational measurement, 3rd* ed. (pp. 221-262). Washington, DC: American Council on Education.

Petersen, N. S., Marco, G. L., & Stewart, E. E. (1982). A test of the adequacy of linear score equating method. In P. W. Holland & D. B. Rubin (Eds.), *Test equating* (pp. 71-135). New York: Academic Press.

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika, 53,* 492-502.

Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement, 14,* 197-201.

Raju, N. S., van der Linden, W. J., & Fleer, P. F. (1995). IRT-based internal measures of differential functioning of items and tests. *Applied Psychological Measurement, 19*(4), 353-368.

Rogers, H. J. (1989). *Item bias investigation with logistic regression.* Unpublished doctoral dissertation, University of Massachusetts, Amherst.

Rogers, H. J., & Swaminathan, H. (1993). A comparison of the logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement, 17,* 105-116.

Roussos, L. A. & Stout, W. F. (1996). Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel Type I error performance. *Journal of Educational Measurement*, 33(2), 215-230.

Rudner, L. M. (1977, April). *An approach to biased item identification using latent trait measurement theory*. Paper presented at the meeting of the American Educational Research Association, New York.

Rudner, L. M., Getson, P. R., & Knight, D. L. (1980). A Monte Carlo comparison of seven biased item detection techniques. *Journal of Educational Measurement, 17*, 1-10.

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores.*Psychometrika Monograph Supplement.

Scheuneman, J. (1979). A new method for assessing bias in test items. *Journal of Educational Measurement, 16*, 143-152.

Shepard, L. A., Camilli, G., & Averill, M. (1981). Comparison of procedures for detecting test-item bias with internal and external ability criteria. *Journal of Educational Statistics, 6*, 317-375.

Sinharay, S., & Holland, P. W. (2006). *The correlation between the scores of a test and an anchor set* (ETS RR-06-04). Princeton, NJ: Educational Testing Service.

Sinharay, S., & Holland, P. W. (2007). Is it necessary to make anchor sets mini-versions of the tests being equated or can some restrictions be relaxed? *Journal of Educational Measurement, 44*(3), 249-275.

Sinharay, S., & Holland, P. W. (2008). Choice of anchor set in equating. *ETS Research Spotlight, 1*, 3-6.

Skorupski, W. P., Jodoin, M. G., Keller, L. A., & Swaminathan, H. (2003, April). *An evaluation of equating procedures for capturing growth.* Paper presented at the meeting of the National Council on Measurement in Education, Chicago, IL.

SPL2K [Computer software]. (2010). Amherst, MA: University of Massachusetts Amherst.

Spray, J., & Miller, T. (1994). *Identifying nonuniform DIF in polytomously scored test items* (American College Testing Research Report Series 94-1). Iowa City, IA: American College Testing Program.

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

Su, Y.-H., & Wang, W.-C. (2005). Efficiency of the Mantel, Generalized Mantel-Haenszel, and logistic discriminant function analysis methods in detecting differential item functioning for polytomous items. *Applied Measurement in Education, 18*(4), 313-350.

Sukin, T. M., Dunn, J. L., Kim, W., & Keller, R. R. (2010, May). *A balancing act: Common Items Nonequivalent Groups (CING) Equating Item Selection*. Paper presented at the meeting of the National Council of Measurement in Education, Denver, CO.

Sukin, T. M., & Keller, L. A. (2008, October). *The effect of deleting anchor items on the classification of examinees.* Paper presented at the Northeastern Educational Research Association, Rocky Hill, CT.

Sukin, T. M., & Keller, L. A. (2009a, October). *Optimal anchor set design: Using generalizability theory to inform item selection.* Paper presented at the Northeastern Educational Research Association, Rocky Hill, CT.

Sukin, T. M., & Keller, L. A. (2009b). *The effect of deleting anchor items on the classification of examinees: An exploration of item flagging criteria* (CEA Report XX). Amherst, MA: Center for Educational Assessment, University of Massachusetts.

Sukin, T. M., & Keller, L. A. (2010, May). *The effect of deleting anchor items on the classification of examinees: An exploration of item flagging criteria.* Paper presented at the American Educational Research Association, Denver, CO.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*, 361-370.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland, and H. Wainer (Eds.), *Differential item functioning* (pp. 67-113). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Thissen, D., & Wainer, H. (1982). Some standard errors in item response theory. *Psychometrika, 47*, 397-412.

Vale, C. D., Maurelli, V. A., Gialluca, K. A., Weiss, D. J., & Ree, M. J. (1981). Methods for linking item parameters (AFHRL-TR-81-10). Brooks Air Force Base TX: U.S. Air Force Human Resources Laboratory.

Wang, W. C., & Wu, C. I. (2004). Gain score in item response theory as an effect size measure. *Educational and Psychological Measurement, 64*(5), 758-780.

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.

Whitely, S. E., & Dawis, R. V. (1976). The influence of test context on item difficulty. *Educational and Psychological Measurement, 36*, 329-337.

Wingersky, M. S., Cook, L. L., & Eignor, D. R. (1987). *Specifying the characteristics of linking items used for item response theory calibration* (Research Report 87-24). Princeton, NJ: Educational Testing Service.

Yang, W. L. (2000, April). *The effects of content homogeneity and equating method on the accuracy of common-item test equating.* Paper presented at the meeting of the American Educational Research Association, New Orleans, Louisiana.

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*, 297-311.

Yen, W. M., & Fitzpatrick, A. R. (2006). Item response theory. In R. Brennan (Ed.), *Educational Measurement*, 4th ed. (pp. 111-153). Westport, CT: Praeger Publishers.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and liker-type (ordinal) item scores* [On-line]. Ottawa, Ontario, Canada: Department of National Defense, Directorate of Human Resources Research and Evaluation. Available: http://www.educ.ubc.ca/faculty/zumbo/DIF/index.html.

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*, 321-324.