STORE, DAVIE, Ph.D. Item Parameter Changes and Equating: An Examination of the Effects of Lack of Item Parameter Invariance on Equating and Score Accuracy for Different Proficiency Levels. (2013)
Directed by Dr. Richard M. Luecht. 266 pp.

The impact of particular types of context effects on actual scores is less understood although there has been some research carried out regarding certain types of context effects under the nonequivalent anchor test (NEAT) design. In addition, the issue of the impact of item context effects on scores has not been investigated extensively when item response theory (IRT) is used to calibrate the items and maintain the score scale. The current study focuses on examining the impact of item parameter changes for anchor test items in a particular IRT equating context. The study specifically examines the impact of different types and magnitudes of item serial position changes as “context effects” on score accuracy and performance-related decisions (e.g., classifying examinees on pass/fail mastery tests or into three or more achievement levels).

The study uses real data from a large-scale testing program to determine plausible levels of item difficulty changes as well as the magnitude of association between serial position changes and item difficulty changes. Those real-data results are then used to specify reasonable conditions of item difficulty changes in a large-scale, IRT-based computer simulation in order to investigate the comparability of different study conditions and Rasch equating methods in terms of adequacy to attaining successful equating within and across test designs.

Results of the study indicate that when items change positions, they become either difficult or easier depending on the direction and magnitude of the change. Apparently,

these changes in difficulty become very notable for low ability examinees in comparison to high ability examinees. Because high ability examinees are already more likely to get most items right, it is more unlikely to notice any changes due to changes in difficulty and /or context effects. To the contrary, with low ability examinees, there is a lot of room to investigate the impact the difficulty of an item has on an examinee; many low ability examinees are already missing many items and therefore decreasing or increasing the difficulty of an item enormously affects the probability of these examinees to respond to the item correctly. Further, examination of bias and root mean squared error statistics showed no differences among Rasch equating methods within testing conditions. However, for similar conditions that only differed in difficulty, results were different.

ITEM PARAMETER CHANGES AND EQUATING: AN EXAMINATION OF THE EFFECTS OF LACK OF ITEM PARAMETER INVARIANCE ON EQUATING AND SCORE ACCURACY FOR DIFFERENT PROFICIENCY LEVELS

by

Davie Store

A Dissertation Submitted to
the Faculty of The Graduate School at
The University of North Carolina at Greensboro
in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Greensboro
2013

Approved by

____
Committee Chair

To my wife Jessie and daughters Tamanda and Tadala
To my brother Francis

APPROVAL PAGE

This dissertation, written by Davie Store, has been approved by the following committee of the Faculty of The Graduate School at The University of North Carolina at Greensboro.

Committee Chair ________________________________

Committee Members ________________________________

____

____

____
Date of Acceptance by Committee

____
Date of Final Oral Examination

ACKNOWLEDGMENTS

I am especially indebted to Dr. Ric Luecht for tirelessly challenging, guiding, and encouraging me throughout this dissertation. Without your help, this dissertation was a far-fetched reality. Thanks for your personal support throughout my graduate study.

Dr. Rick Morgan’s and Dr. Terry Ackerman’s attention to details and practical ways of approaching psychometric issues inspired and guided me. I appreciate all you have taught me. I also owe my success to Dr. Micheline Chalhoub-Deville and Dr. Sarah Berenson for their personal support and opportunities for professional experience.

Thanks to Dr. Mike Nering and the entire psychometric team at Measured Progress for the support you offered. You are a dream team and you continue to provide environments that are conducive to nurturing the growth of developing psychometricians.

I would like to acknowledge my colleagues, friends, and professors from Educational Research Methodology Department at UNCG including Drs. Devdass Sunnassee, John Willse, and Bob Henson, whose insights were essential pieces of this dissertation. Special thanks to Steve and Missy Hinerman for being friends who provide opportunities for making things possible.

As is the case with all my achievements, this one owes much to those who raised me. This success is because of the sacrifices and love of my parents Theressa and Amos Daniel Store, my uncle and aunt Cullen and Martha Ndala, and the village that raised me. Also, Tadala, Tamanda and Jessie, thanks for the time and support you gave me. I am what I am because of your support.

Finally, I acknowledge the teachers who inspired me. I never saw this one coming as I sat on the rocks and the banks of Mbombwe River.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

CHAPTER I

INTRODUCTION

The field of testing is currently going through significant changes with increased emphasis on student and educator accountability for learning, a proliferation of professional certification and licensure tests, and increased interest in levering new technologies for assessment needs such as adaptive testing and the expanded inclusion of performance-based and technology-enhanced item types. Nonetheless, many testing programs continue to use some number of “fixed” test forms—that is, forms where items are pre-assigned to specific serial positions in the test and all examinees assigned a particular form see the same items in the same presentation order.

The use of multiple test forms within and across test administrations with common items across the test forms, a trademark of many standardized tests, is usually for two reasons: (a) test and item security and (b) to facilitate pretesting new items. Within a test administration window, having multiple forms prevents certain types of cheating and collaboration. For example, randomly assigning different forms to examinees within the same test center or classroom reduces the likelihood of copying because examinees in close proximity to one another have different forms. Similarly, having multiple forms lessens the risk of collaboration among examinees who might otherwise be induced to conspire to memorize and share items and the supposed answers—especially if forms are active for more than one day or span time zones in the case of national or international examinations.

However, the presence of multiple forms introduces a number of complexities related to the comparability of scores for examinees taking the different forms. The *invariance* principle of measurement states that examinees ought to be indifferent as to which particular form of a test they take (Kolen & Brennan, 2004; Lord, 1980). The purpose of statistical equating is to ensure the comparability of scores—to realize the invariance principle even though test forms may differ in difficulty and other statistical characteristics. If equating works, we should have a *common score scale* regardless of the test form taken —that is, interchangeable scores.

There are three ways in which statistical comparability of scores can be achieved: (1) randomly assigning test forms where examinees are assumed to be sampled from a common population; (2) using common items to link forms where randomly equivalent sampling cannot be assumed; or (3) having the same examinees take two or more test forms (Kolen & Brennan, 2004). None of these approaches is foolproof. Many issues, including test administration scheduling and logistical limitations, can contaminate even the best random sampling designs. Maintaining common items—the topic of this dissertation—can be susceptible to item exposure/disclosure, learning/opportunity to learn changes, curricular, content or “factual knowledge” changes over time, cheating, and a myriad of other factors that alter the apparent statistical characteristics of the linking items. Finally, many practical and logistic issues such as increased test administration costs, motivation effects, and natural maturation/growth often preclude using common persons to establish comparability.

This dissertation is about the second issue when we cannot reasonably support an assumption of randomly equivalent examinee groups taking the test forms and common items link the forms. The equating assumption is that the common linking items must have the same operating characteristics regardless of the form on which they are used. If that assumption holds, we can attribute any differences in performance on those common items to proficiency differences in the examinee groups taking each form. Some research suggests that context effects such as serial position of the common items matters (Cook & Petersen, 1987). Cizek (1994) has suggested that more subtle changes such as reordering the response options on common multiple-choice linking can alter the statistical characteristics and affect the accuracy of the equating process. Certainly, factors such as changes in educational curricula over time and item/test-form disclosure policies affecting the common items would also affect the equating in perhaps unknown ways. This dissertation considers the specific factor of item context effects (e.g., changes in statistical item difficulty and/or discrimination) when the linking items are forced in new positions on different test forms.

An example may help. Some testing programs randomly assign items to fixed-location pretest blocks or “slots” during the initial tryout period for those items. The same items then migrate to other positions or blocks in the test forms for operational use (i.e., as scored items) Figure 1.1 illustrates this positional shift from pretest to operational status. If linking items are shared across the forms, we can use those common items to equate the test forms to one another or to an underlying metric or scale. Conventionally, we refer to the linking or common items as the “anchor test” (AT). As Kolen and Brennan (2004) note, using linking AT items is the only way to equate score scales for two or more examinee groups that are potentially from different populations (e.g., examinees taking the test at different times within the year or across years). We refer to those groups as non-equivalent (NE), implying that we simply cannot assume the population proficiency score distributions to be randomly equivalent. Using the common AT items in the context of NE groups leads to what is referred to as the NEAT (non-equivalent anchor test) design. As noted above, if the AT item difficulty and/or discrimination statistics change over time, it is typical to assume that those statistical differences result from solely the differences in the examinee score proficiency distributions. However, if there are testing context effects (e.g., changes in item serial positions) that result in direct or indirect effects on the statistical characteristics of the AT items could contaminate the equating assumptions of the NEAT design and lead to biased equating and associated scoring/decision-making errors.

Although there has been some research carried out regarding certain types context effects under the NEAT design, the impact of particular types of context effects on actual scores is less understood. Furthermore and highly germane to this study, that issue of the impact of item context effects on scores has not been extensively investigated when item response theory (IRT) is used to calibrate the items and maintain the score scale. The current study focuses on examining the impact of item parameter changes for AT items in a particular IRT equating context and specifically examines the impact of different types and magnitudes of item serial position changes as “context effects” on score accuracy and performance-related decisions (e.g., classifying examinees on pass/fail mastery tests or into three or more achievement levels).

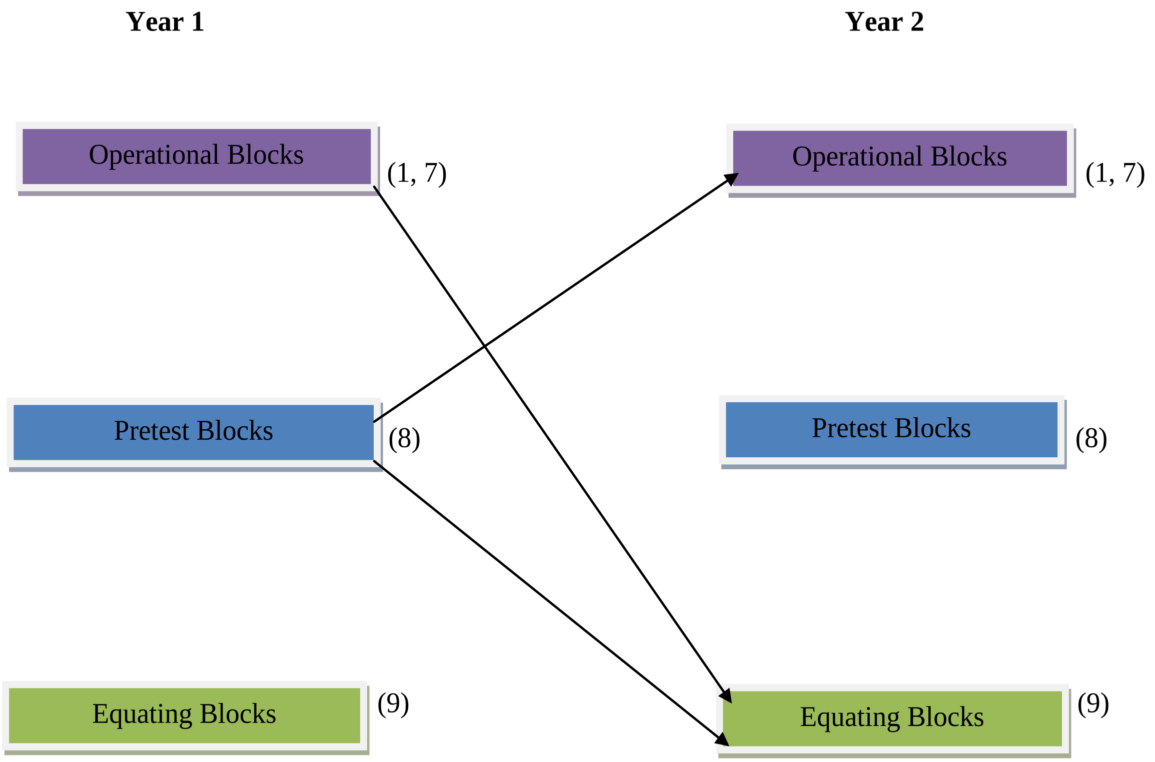


Figure 1.1. A schematic diagram showing how items can change from one year/administration to another when fixed blocks for pretest and equating are used.

The IRT calibration and equating framework chosen for this study mimics the type of operational equating design that many state departments of education in the United States employ for their kindergarten to grade 12 (i.e., K-12) end-of-grade and end-of-course examination programs. That framework involves calibrating items using a Rasch IRT model to a common item bank scale—typically denoted by the Greek letter “theta” (θ). Proficiency-level cut scores are maintained on the θ metric and used to classify examinees into achievement-level categories such as “basic,” “proficient,” and “advanced.” In turn, those reporting categories help chart educational progress of the students over time as well as for accountability purposes. The AT items are calibrated to the common θ metric underlying the item bank. Those items are then reused on future test forms to provide the required equating links between all new test forms and the item bank scale. Much like any NEAT equating design, the IRT Rasch equating process uses the characteristics of the AT items as the sole basis for statistically adjusting each new test form calibration so that all estimated scores for all examinees are on the same item-bank scale, θ.

The study uses real data from a large-scale testing program to determine plausible levels of item difficulty changes as well as the magnitude of association between serial position changes and item difficulty changes. Those real-data results are then used to specify reasonable conditions of item difficulty changes in a large-scale, IRT-based computer simulation.

Statement of the Problem

The issue of context effects in equating came to national prominence with what has been called the National Assessment of Educational Progress (NAEP) “reading anomaly” (Kolen & Brennan, 2004). The NAEP anomaly resulted in a rather steep score drop between 1984 and 1986 in estimated reading proficiency scores. That drop was later attributed to changes in the order and context in which anchor test items appeared on the NAEP forms (Zwick, 1991). It also served as a wake-up call to the psychometric community; context effects may not be ignorable, especially for AT items.

In short, the NAEP reading anomaly provided a concrete reason to begin to investigate the problems that could become manifest due to inconsistencies in the presentation order of anchor test items. Before the NAEP reading anomaly report, most context-effect studies focused on full-length tests and discussed the effects of item arrangement on examinee performance under both speeded and unspeeded test conditions (Dorans & Lawrence, 1990; Eignor & Stocking, 1986; Harris, 1991; Leary & Dorans, 1985). The emphasis in many of these research studies was to detect and estimate the size and consequences of context effects. Kingston and Dorans (1984), Leary and Dorans (1985), and Davey and Lee (2010), among others, describe context effects as changes that occur when examinees' item responses are directly or indirectly affected by factors other than the primary trait or construct being intentionally measured by the test. These factors include the location of an item within a test (Davey & Lee, 2010; Hill, 2008; Meyers, Miller, & Way, 2009; Whitely & Dawis, 1976; Yen, 1980), wording, content, format (Kingston & Dorans, 1984; Zwick, 1991) and specific features of other items that surround it (Davis & Ferdous, 2005; Haladyna, 1992).

Although the obvious solution would seem to be NOT to change the context of items—especially the AT items, that is seldom a feasible solution. The proliferation of testing over the past few decades, increasing stakes of the examinations for test takers educators or others, and testing policies such as item disclosure and item tryout policies have made it difficult for testing organizations to maintain consistent contexts such as similar item positions across test forms. In short, the pragmatic reality of test development costs and the logistical design of test forms often make it difficult to hold constant the relevant context effects. Many research studies have been conducted using classical test theory and demonstrate the impact of those effects (Davey & Lee, 2010; Hill, 2008; Pommerich & Harris, 2003; Whitely & Dawis, 1976). Other research has been carried out using IRT Rasch models (Kingston & Dorans, 1984; Meyers et al., 2009; Yen, 1980; Zwick, 1991).

Item response theory (IRT; Lord, 1980) changed some of the thinking about items and test forms as presenting a unique context for each examinee. Under IRT, items calibrated to a common metric can be reused on new test forms and used to link those forms to that same, underlying metric, θ. There has been IRT-related research that considers test context effects and associated effects, however, it tended to focus on the potential changes in the item parameter estimates—that is, challenges to the “invariance” of the IRT item statistics over time and contexts (Hambleton & Swaminathan, 1985; Baker, 1985; Meyer et al., 2009). Invariance in the IRT context implies that we can obtain an estimate of every examinee’s score on the common θ metric regardless of which test form is administered as long as we know the IRT-calibrated item parameters. For example, this invariance principle is central to computerized adaptive testing (CAT) where examinees are intentionally administered test forms targeted to their apparent proficiency level. We likewise assume under IRT that we can estimate the item parameters using examinee samples that differ in proficiency, as long as we know their proficiency scores. Some research has shown that various threats to item-parameter invariance can have nontrivial consequences on scores and related decisions (Hill, 2008; Meyers et al., 2009; Wise, Chia, & Park, 1989).

Some context-effect studies have highlighted the potential problems (Davey & Lee 2010). But, little has been done to offer concrete solutions. For example, there is a gap in the measurement research literature as to how item position-related context effects interact with other characteristics of the test and examinee population (e.g., the density of measurement information relative to the density of examinee proficiency) and how different IRT equating methods might contend with those effects. There is also very little research that specifically addresses the potential causal relationship between the magnitude of item parameter shifts due to serial position changes on the accuracy of examinees' proficiency scores. This research therefore seeks to fill those gaps and specifically evaluate which of several IRT Rasch-based equating strategies might help mitigate the problems.

Purpose and Rationale

It is impossible to address all viable issues regarding test context effects in one study. Nevertheless, some research has indicated that the relevant scoring-related issues pertaining changes in test contexts first become manifest as changes in the item parameter estimates (Leary & Dorans, 1985). That is, because each examinee's estimated proficiency score is a function of the item parameter estimates used, those score estimates must change if the item parameters change (Wells, Subkoviak, & Serlin 2002). This study therefore starts by focusing on serial position changes as one example out of many test contexts effects that threatens the stability of item parameter estimates. An exploratory analysis of real reading and mathematics test items provides some basis in reality for postulating a statistical relationship between serial item position changes and item difficulty changes. This part of the study also builds on empirical research by Wise et al. (1989) to look at the potential interaction between context effects and examinees at different proficiency levels.

As alluded to earlier, in an ideal world, we would simply hold item context effects constant across all test forms and theoretically eliminate any potential problems. Real test development does not occur in an "ideal world." Kolen and Brennan (2004) have cautioned that equating is challenging because there are so many diverse practical issues and few generalizable solutions that work in every case. There are practical realities, logistical and economic limitations, and human judgments that often go far beyond the rather pure mathematics that underlie statistical models and optimal sampling designs with convenient assumptions. For example, test development issues play an important role in equating and may limit our capability to ever even approximate the "ideal world." As Mislevy (1992) observes,

> test construction and equating are inseparable. When they are applied in concert, equated scores from parallel test forms provide virtually exchangeable evidence about students' behavior on the same general domain of tasks, under the same specified standardized conditions. When equating works, it is because of the way tests are constructed . . . (p. 37)

Since we cannot develop the perfect tests from an equating perspective or hope to control all of the factors that impact scores—despite our best efforts at standardization and sampling—we need to accept a certain inevitable amount of inaccuracy in our results. The second part of this dissertation, however, asks whether certain conditions of test design exacerbate or minimize the problem. For example, how does the density of measurement information (i.e., test information in the IRT context) for the anchor test relative to the full test impact the results? How does the density of measurement information relative to the examinee proficiency score distribution influence the impact of AT item context effects? How does test length (full test and AT) impact results? How does the magnitude of the difference in the non-equivalent group score distributions impact the results? Finally, are there different equating strategies that work better in some of these cases but not others? The second part of this dissertation investigates these and other questions by considering the interactions between a number of specifically manipulated test design and sampling conditions. Ultimately, this study is intended to demonstrate which, if any, sets of conditions might render ANY equating to be ill-advised.

The second part of this dissertation uses a large-scale IRT-based computer simulation. The practical value of this simulation is two-fold. First, it allows for direct manipulation of the item characteristics employed on each test form, including the anchor test items. Item context effects are specifically added to the simulations in known amounts and under specific conditions. Second, because the IRT-based simulations base the data generation on known θ scores and IRT item parameters, it is relatively straightforward to evaluate the accuracy of estimates relative to the "true" values of the parameter using residual-based statistics.

Research Questions

In reference to the above considerations, this research study addresses the following four research questions and sub questions:

1. How does the magnitude and direction of item difficulty changes and conditional probability changes relate to serial item position changes at different proficiency levels on the ability scale?
2. How do different study conditions and Rasch equating methods compare in terms of adequacy to attaining successful equating?
    - 2.1. Which study conditions lead to worst/best equating?
    - 2.2. How do the five equating methods differ in terms of equating bias and RMSE?
    - 2.3. Which method(s) of equating result in somewhat adequate equating, if any, under worst study conditions, among the five equating methods?
    - 2.4. Are the findings discussed above consistent over various test lengths?
3. Is there any advantage (precision gain) in using stability equating over fixed number of anchor items equating?
    - 3.1. How do similar equating methods that only differ in treatment of anchor items (fixed versus stabilized) compare?
    - 3.2. Does pruning of unstable equating items have more effect on equating for shorter anchor item tests than for longer anchor item tests?
4. How does the magnitude of item difficulty changes affect decision accuracy at different proficiency levels of the score scale?
    - 4.1. Which study conditions result in worst/best classification rates for both lower ability and high ability examinees?

4.2. How are classification rates for similar study conditions affected by proficiency level?

4.3. Under worst study conditions, which method(s) of equating result in better classification rates?

4.4. In general, which method(s) of equating result in somewhat adequate equating?

4.5. Are the effects consistent over various test lengths?

The first research question was addressed using data from two large-scale testing programs that allow items to change position as they migrate between pretest, operational and equating status. Alternatively, simulations were used to answer the other three research questions. Model-based simulations allow evaluation of more conditions to provide a direct way to compare estimates to the known (generated) parameters. In the present context, the primary parameters of interest are IRT examinee scores (denoted as θ or “theta”).

Definition of Terms

When items change positions between test forms, serial position context effects are expected. As has been discussed already, these effects show up as changes in item parameters. Items may move towards the beginning of the test from one administration to another. Such movement is defined as negative change in this study. Alternatively, items may move towards the end of the test from one administration to another, a condition that is termed positive change. This definition of position change is consistent with Meyers et al. (2009). The effect of such item movements on item difficulty is what this study seeks to discuss.

Further, if item parameters are affected, examinees' probabilities of getting a correct response may also be affected in two ways. The first case is when the probability of getting the correct response becomes lower than the probability during an item's initial administration. This case is defined as negative probability change (difference), which entails that for a given examinee, an item is harder than it was when initially administered. Alternatively, positive probability change (difference) is characterized by having a higher probability when an item is readministered than that of initial administration. This shows that an item is easier.

However, because different examinee groups may experience varying amounts of difficulties due to different item arrangements (Wise et al., 1989), examining the effects of lack of item invariance for different proficiency levels becomes a requirement. In real testing situations, a prescribed amount of ability score demarcates the boundaries between proficiency levels. These demarcations are called cuts or cut points. For this study, cut 1marks the boundary between substantially below proficient and partially proficient examinees, cut 2 draws the line between partially proficient and proficient examinees while cut 3 marks the boundary between proficient and advanced examinees.

Finally, the success of any equating process is ultimately evaluated by the adequacy through which practical issues have effectively been handled (Kolen & Brennan, 2004). When all or most practical limitations, constraints, and complications have been addressed, we may be able to conclude that the equating is at least “adequate” (i.e., the best practice possible, under the circumstances).

Delimitations

While this research presents typical concerns about threats to item parameter-invariance and their impacts on equating and that the findings can generalize to almost all situations where IRT models are used, the present conditions discussed herein determine the scope of applicability of the findings. Users of this research study are therefore encouraged to analyze how their needs align with those documented in this study. For example, examinees in the testing programs used in this study take tests within prescribed amounts of time, without any knowledge of field-test or operational items, which makes the examinees to be equally motivated to take the field and operational test items. This study therefore assumes that if item position remains the same, item characteristics from field-testing will be similar to operational statistics. In light of these circumstances, applying the findings of this study to other situations outside these prescribed conditions, e.g., where field-test items are disclosed to examinees, tests are untimed, classical test equatings are conducted and multidimensionality exists may be incongruous with the purposes of this research.

CHAPTER II

REVIEW OF THE LITERATURE

> In operational testing programs using IRT, model advantages must often be weighed against concerns over threats to item parameter invariance. One place where this occurs is the position of items in a test from one use to the next. The conservative view is that item sets used for test equating remain in identical or very similar positions within the test from one use to the next. Of course, this requirement can be limiting and difficult to sustain over time. (Meyers et al., 2009, p. 39)

As Zwick (1991) points out, changes in item order, context, and time allocated to complete the common items may at times seem incautious because of attempts to maintain consistency with current practice while embracing the optimistic views that prevail on the robustness of item response theory. Overdependence on the assumption that these changes will not have serious consequential effects on item parameters seems unrealistic.

Background to Item Response Theory

Currently, many testing practices depend on IRT. For decades, IRT has been the building block for many issues in testing that evolved from classical test theory (CTT). Hambleton and Swaminathan (1984) discuss five shortcomings of classical test theory that propelled the emergence of item response theory. Among other shortcomings, the magnitudes of commonly used item statistics such as item difficulty and discrimination are dependent on the sample of examinees used. Again, with CTT the influence of average and range of ability of the examinees on item statistics is inevitable. Ultimately, classical test theory statistics are not applicable to populations of examinees that are different to the sample of examinees in which the item statistics were obtained and are useful only in item selection when constructing tests for examinee groups that are similar.

Additionally, classical test theory provides no basis for determining how an examinee might perform when confronted with a test item. As Hambleton and Swaminathan (1985) illustrate,

> having an estimate of the probability that an examinee will answer a particular question correctly is of considerable value when adapting a test to match the examinee's ability level. Such information is necessary, for example, if a test designer desires to predict test score characteristic in one or more populations of examinees or to design tests with particular characteristics for certain populations of examinees. (Hambleton & Swaminathan, 1985, p. 3)

With these classical test theory inadequacies in mind and other factors like failure to provide satisfactory solutions to many testing problems such as test designs, differential item analysis (DIF) and test equating (Hambleton & Swaminathan, 1985), among others, psychometricians have resolved to developing theories of mental measurement that are tailored to overcome such inadequacies and are conversant with today's testing needs. The emergence of item response theory is a product of such innovative works by psychometricians to reach a common goal of making inferences about unobservable examinee traits from test responses. Mills, Potenza, Fremer, and Ward (2002) have indicated that most researchers in early work (Birnbaum, 1968; Ferguson, 1942; Lawley, 1943) refer this approach as *latent trait theory*. Lord (1952) used the term *item characteristic curve theory* and the theory is now *item response theory* (Lord, 1980).

More definitively, Baker (1985), van der Linden and Hambleton (1997), Folk and Smith (2002), DeMars (2010), and de Ayala (2009) have comprehensively discussed the basics and applications of item response theory and reiterate that this theory supposes that examinee traits or abilities can be inferred from examinee performance on a test. DeAyala (2009) defines item response theory as, effectively, a system of models that defines one way of establishing the correspondence between latent variables and their manifestations. More precisely, Hambleton and Swaminathan (1984) describe that an item response model specifies a relationship between observable examinee test performance and the unobservable traits or abilities that underlie performance on the test. Baker (1985) adds that a reasonable assumption for item response theory is that each examinee responding to a test item possesses some amount of the underlying ability, also called theta. Different examinees will have different ability levels because of the amount of theta they have which ultimately translates to different probabilities for responding to an item given the amount of theta. Wilson (2005) provides a good illustration of construct-response where he argues that the idea of causality is just an assumption and therefore, confirmation of directionality to reveal the nature of the relationship should follow from research.

However, for IRT models to be applicable, three main assumptions should be satisfied. One of these assumptions is that the test itself should be unidimensional. DeMars (2010) clarifies,

> whenever only a single score is reported for a test, there is an implicit assumption that the item shares a common primary construct. Unidimensionality means that the model has a single theta for each examinee, and any other factors affecting the

> item responses are treated as random error or nuisance dimensions unique to that item and not shared by other items. (p. 38)

The second assumption of item response theory is local independence, which simply implies that there is no statistical relationship between/among examinees' responses to different items in a test. Precisely, examinee's performance on one item must not provide an advantage or disadvantage to his/her performance to other items on the test (Hambleton & Swaminathan, 1984). While DeMars (2010) recommends the importance of having the items correlating in the sample as a whole in order to show the unidimensionality aspect, she quickly points out that this should not be the case after controlling for theta. As will be pointed out in later sections from studies conducted by Hambleton and Traub (1974), local independence is usually under threat when presentation of items on a test affects test performance.

Finally, in order to use item response theory, correct model specification assumption is a necessary requirement. Currently, the one parameter logistic (1 PL) model, also referred to as Rasch model, is the model that is commonly being used by many test developers. The Rasch model,

$$P_i(\theta) = [1 + exp(b_i - \theta)]^{-1},$$

simply expresses the relationship between θ, the proficiency score (latent trait) and b_i, the item difficulty parameter. Because of its simplistic nature, it allows test developers to explain the relationship between ability and difficulty without any confounding interpretations that occur due to the existence of other parameters in the model. However, as users/proponents of the three parameter logistic (3 PL) model will argue that it is important to acknowledge the fact that some examinees at the lower end of the ability scale may be expected to have a high probability of providing a correct response. De Ayala (2009), express that two concerns call for the need for the development of the three-parameter model. The first concern, which is also modeled by the two-parameter (2PL) model, addresses the question of finding the probability of a response of one on an item when an examinee responds consistently with his/her location on theta. The second concern addresses the question of finding the probability of the response of 1 on an item due to chance alone. For these reasons, users/proponents of the 3PL model sacrifice ease of interpretation that the Rasch model offers with better fit that the 3 PL brings about. The 3 PL is expressed as follows:

$$Prob(u_i = 1|\theta) \equiv P_i(\theta) = c_i + (1 - c_i)\{1 + exp[-a_i(\theta - b_i)]\}^{-1},$$

which mathematically models the probability of a correct response to a dichotomously scored item i, given θ, the examinee’s latent proficiency score. Item characteristics are represented by a_i, an item discrimination parameter related to the slope of the probability function, b_i, an item difficulty or location parameter, and c_i, a lower asymptote parameter associated with noisy response patterns exhibited by lower-performing examinees, possibly due to guessing.

The incorporation of the pseudo guessing parameters could not have occurred without some controversies. As has been pointed out already, the pseudo guessing parameter reflects that some examinees with infinitely low locations may obtain a response of 1 when according to the two-parameter model they should not. De Ayala's insightful observations led to the following statement;

> these responses are a manifestation of the interaction between person and item characteristics (including item format). In the case of proficiency instruments, person characteristics include not only a person's theta, but also her test wiseness and risk-taking tendencies. These last two factors are tangential latent personal variables. Therefore, although c_i is considered to be an item parameter, it may be more reflective of a person characteristic (i.e., another person parameter) rather than an item characteristic or, at least, an interaction between person and item characteristics. (De Ayala, 2009, p. 126)

Item Response Theory and Testing

One of the most practical uses of item response theory in the testing field is the use of item invariance principle. Baker (1985) describes that this principle entails that examinee's ability is invariant with respect to the items used to determine it (item parameters are invariant across different examinee samples) and is based on the conditions that all items measure the same underlying latent trait and that all item parameters are in common metric. Hambleton and Swaminathan (1984) precisely express that ability estimation independent of the particular choice and number of items represents one of the major advantages of item response models. The invariance assumption is one of the rock solid foundations on which item response theory applications rest and makes it possible for reasonable ability comparisons between different examinees in terms of their performance based on item parameters despite the sample used to calibrate the items. Rubin and Mott (1984) emphasize that the invariance principle allows test developers to gather item statistics in one occasion and use the information subsequently to compile tests having predetermined characteristics. Meyers et al. (2009) extend this point by clarifying that the item-parameter invariance principle has allowed researchers to apply item response theory to other areas such as computer adaptive testing and test pre-equating. It can therefore be said that the practical implication of the invariance principle is that a test located anywhere on the theta scale can be used to estimate an examinee's ability. However, when the invariance principle fails to hold, it results in differential change of item parameters over subsequent testing occasions. As DeMars (2010) observes, these changes in parameter estimates might be due to a shift in instructional emphasis, disclosure of the items by previous test takers, or changes in the construct over time. Wells et al. (2002) call this differential change *item-parameter drift.* They elaborate that item parameter drift from whatever cause poses a threat to measurement applications that require a stable scale and that because under item response theory, an examinee's ability is a function of the item parameters. Therefore, ability estimates for examinees will change, if the item parameters change.

Hambleton and Swaminathan (1984) provide three primary advantages of item response theory, which in my view, are the benefits of using item response theory that are reflected by modern testing practices. First, most testing institutions have developed large pools of items all measuring the same trait to make use of the invariance principle where estimation of examinee ability is independent of particular items. Second and conversely, testing institutions have made use of large samples of examinees to calibrate items that result in stable item characteristics and are not dependent on the samples used in the calibration. Finally, the provision of precision parameters for all examinee ability estimates in item response theory cannot be understated.

However, with all the documented beneficial practical implications to testing practices that have emerged with IRT, and more specifically with the invariance principle, a number of threats to the parameter invariance principle have been established, especially at item level. Among the many threats of item invariance principle are item location effects (Kingston & Dorans, 1984), item order effects (Hambleton & Traub, 1974), instructional effects, variable sample sizes and other sources of item parameter change that are informally known in item response theory applications (Meyers et al., 2009).

Calibrations and IRT Equating Methods

Calibration refers to a process of determining the statistical item characteristics using IRT. IRT calibrations are used to determine an underlying metric or scale that can be simultaneously used to locate examinees or items. As has been pointed out already, equating refers to the statistical process of adjusting a particular IRT calibration to the base (previously calibrated item-bank) scale. As Petersen (2008) states, the need to equate test scores is a result of the test developer's inability to construct multiple forms of a test that are strictly parallel. Therefore, the process of equating is an attempt to fine-tune the test construction process. Equating or linking a calibration is tantamount to maintaining the continuity of the base scale and proficiency standards over time.

Anchor items are crucial for successful equating. As Kolen and Brennan (2004) discuss, when a NEAT design is used, anchor item sets should be built to the same specifications, proportionality, as the total test. Stated differently, anchor tests should be miniature versions or minitests of the tests being equated (Sinharay & Holland, 2006), or as Angoff (1968) and Budescu (1985) recommend, an anchor test that is a parallel miniature of the operational forms. This means that, there should be enough number of anchor items on the test to represent the test content covered in the whole test. However, Sinharay and Holland (2006) have argued that requiring an anchor test to mimic the statistical characteristics of the total test may be too restrictive as anchor tests with a spread of item difficulties less than that of a total test seem to perform as well as a minitest with respect to equating bias and equating standard error. Nevertheless, to this day many test developers abide to the requirement that anchor tests be representative of the total tests with respect to content for justification from the perspective of content validity.

In separate studies, Budescu (1985) and Wingersky, Cook, and Eignor (1987) have indicated that larger numbers of common items lead to less random equating error while Petersen, Cook, and Stocking (1983) pointed out that when few anchor items are used, equating problems arise. Therefore, for attainment of adequate equating, Kolen and Brennan (2004) suggest that a common item set should be at least 20% of the test length when a test contains 40 or more test items.

In addition, anchor items should not function differently in the old and new forms. In order to achieve this, anchor items should be administered in almost the same positions in the old and new forms (Cook & Petersen, 1987). Also, the response alternative order should not be changed (Cizek, 1994).

The process of equating requires that the following five conditions be satisfied; (a) same construct, (b) equal reliability, (c) symmetry principle, (d) equity principle

(Lord, 1980), and (e) population invariance principle. The first condition entails that tests must be measuring the same characteristic, latent trait, ability, skill, or construct while the second condition ensures that score comparisons are from almost equally reliable tests. The symmetry principle ensures that score transformation must be invertible. Stated in other words, once score transformation from one form to the other form's equivalence takes place, the reverse process should translate to the original score. As for the equity principle, it highlights that it must be a matter of indifference for applicants at every given ability level whether they are to take one form or the other whereas the population invariance principle emphasizes that score transformation should be the same regardless of the group from which it is derived.

IRT equating involves selecting a design, placing item parameters on a common scale and using the relationship between abilities and true scores on the two test forms that require equating to establish the raw-to-scale relationship. For this study, a non-equivalent anchor-item test design is embraced. In this design, two groups of examinees, that are not randomly equivalent, take different forms of a test with a common set of items. It is therefore important for the common items to reflect the compositions of the two tests that require equating both statistically and in content composition since the common set of items is a reference point through which group differences are determined.

Cook and Eignor (1991) have expressed that the NEAT design is probably the most difficult to execute technically because the quality of the equating depends on the similarity of the groups taking the new and old forms of the test, the parallelism of the two tests to be equated, and the quality of the anchor test. The central task in equating using this design is to separate group differences from form differences (Kolen & Brennan, 2004). The use of IRT equating with a NEAT design may be the best method to use when nonrandom groups of examinees who differ in ability take tests with differing difficulties. Second, because of the invariance principle, IRT equating provides conversions that are independent of the group or groups used to obtain them.

In the NEAT design, two examinee groups, *P* and *Q* take two test forms X and Y that have a set of common items. Because the two test forms are administered to two examinee groups that are assumed to be non-equivalent, the set of common items should be proportionally representative of the total test forms in content and statistical characteristics in order to reflect group differences accurately (Kolen & Brennan, 2004). Similarly, as has been brought up in the discussion on context effects, Cook and Petersen (1987) suggest that the common items should occupy same positions in the old and new forms and that the common items should be the same (e.g., no wording changes or rearrangement of alternatives).

In conclusion, as Mislevy (1992) highlighted, the process of equating cannot be detached from test development processes. Kolen and Brennan (2004) elaborate that if problems exist with test construction, no amount of statistical manipulation can lead to successful equating. It is important that test developers pay attention to practical issues and apply informed decisions to accomplish adequate equating.

Historical Background on Context Effects

To a large extent, item location and order effects have been discussed broadly in the literature from the early 1950s (Mollenkopf, 1950) and up to the present day, the debate on item arrangement still goes on. Leary and Dorans (1985) provided a historical perspective on the implications of altering the context in which test items appear and discuss three main themes on what seems to be the driving force behind active research in this area which has clocked over six decades. They mention that literature has produced evidence of context effects, but has not demonstrated that the effects are so strong as to invalidate test theory or practice that is dependent on an assumption of item-parameter invariance. Leary and Dorans (1985) have mapped out three definitive periods in context effects research. They further elaborate that many of the salient and common features of context effects research are a function of the practical psychometric concerns of the period of concern. For example, initial attempts to use new technology and resources to gain a better understanding of tests and their use motivated the earliest literature on context effects, extending from 1950 to the late 1960s.

> The present concern regarding the moving of test items or sections to accomplish pre-equating or to develop adaptive tests has shifted the emphasis to the effects on item parameters that might result from changing item orders. To draw conclusions about the effects of item rearrangement on test performance for the purpose of answering the most recent questions, analysis of common characteristics of the research, across these three broad time periods, must be considered. (Leary & Dorans, 1985, p. 387)

There is enormous documentation of item-arrangement research studies in the psychometric literature focusing on a wide range of topics. Some studies have focused on factors related to the main effect of item arrangement on examinee performance for both speeded and power tests. These studies have discussed issues like random section scrambling (Mollenkopf, 1950) and random item scrambling (Monk & Stallings, 1970; Hambleton & Traub, 1974). Other studies have focused on easy/hard vs. hard/easy vs. random (Klein, 1981; Sax & Cromack, 1966). While some studies have focused on scrambling sections and arranging difficulties of items within sections (Brenner, 1964; Flaugher et al., 1968), other studies have focused on context manipulation (Huck & Bowers, 1972; Sax & Carr, 1962). Findings from these early studies indicate that item arrangement affects the performance of examinees on a test and in addition, speededness mediates the effects of item arrangement on performance. For example, if a speed test has items arranged from difficult to easy items, low ability examinees may not be able to reach the easy items as they might spend a lot of time working on the difficult items. Because of this, examinees spend much time on the items that will not give enough information on their ability.

Leary and Dorans (1985) clarified that it was the finding that not only the main effect of item arrangement on examinee performance played a role in explaining item difficulty that prompted researchers to include interactions with biological and psychological characteristics. Some studies done during this period (from the 60s to 80s) addressed the issues of item order and anxiety (Berger et al., 1969; Marso, 1970; Smouse & Munz, 1968, 1969; Towle & Merill 1975). Other studies focused on item order, anxiety, and sex of the examinees (Hambleton & Traub, 1974); item order, anxiety, knowledge of order and sex (Plake et al., 1981, 1982); item order and achievement level

(Klossner & Gellman, 1973). In summary, research conducted in this era aimed at investigating whether item arrangement had a different effect on performance for males and females or whether item arrangement has an effect on examinee performance depending on how anxious the examinee is or whether item arrangement has such adverse effects on performance depending on examinee proficiency level. The findings from these studies are not definitive as they vary from study to study. For example, earlier studies by Smouse and Munz (1969) indicated the existence of item order by performance anxiety-type interaction effects but replications of such studies by Towle, Merill, Berger et al., proved fruitless (Leary & Dorans, 1985). Therefore, the effects of item arrangement on test performance with respect to test anxiety, achievement level and examinee sex have not been precisely established due to these erratic findings from sample to sample. However, Leary and Dorans (1985) have indicated that the most definitive result in this area of research is that hard to easy arrangement of items yield lower scores on speeded tests than easy-to-hard arrangements.

As has already been mentioned, the driving force behind investigating effects of item location keeps changing due to the changing needs of the testing industry over time. More recently, motivated by the impact of test disclosure legislation on data collection designs for the equating of new test forms, there have been discussions on the effect of repositioning intact test sections. Leary and Dorans (1985) observe that embracing section pre-equating and item response theory pre-equating methods has come along with two major problems. First, item parameters or estimated section difficulty with respect to test taker's performance may not be constant during operational use. Second, pre-equated sections do not introduce practice effects as additional factors affecting test performance.

While most studies discussed in the preceding sections have investigated effects of item arrangements for the whole test, the late 70s saw the move to discuss issues that are pertinent to anchor items and specifically investigating the stability of the item parameter invariance associated with models of item response theory. Whitely and Dawis (1976) investigated context effects on classical item difficulties (*p*-values) and Rasch item difficulty for a verbal analogies test comprised of 60 items of which 15 items were common items placed in seven different test forms which were tested on seven different samples of examinees with each sample having about 200 examinees. Using analysis of variance, they found out that nine of the fifteen common items had statistically significant differences in *p*-values and had similar patterns in Rasch item difficulty.

In line with Whitely and Dawis (1976), Yen (1980) conducted a study to investigate the causes and importance of item context effects for the three-parameter logistic and the Rasch models. Yen used three different classes of discrimination and difficulties based on pretest item estimates to create seven reading forms and seven mathematics forms where six forms of each (math or reading) were randomly tested and the seventh form was tested two weeks later. Yen found that changes in item arrangements decreased the stability of item difficulty for both models. He concluded that there were item context effects especially for the discrimination parameters even though discrimination was highly affected by the number of items used for calibration. As for difficulty, Yen found out that items in reading passages were generally more difficult when they appeared towards the end of the test.

Because the central focus of item response theory is to estimate true scores from observed scores, Kingston and Dorans (1984) carried out an investigation to assess within test context effects on item response theory true score equating. Kingston and Dorans used item statistics from a 3PL model to investigate the effects of item position on examinee performance for two forms of the 1980 GRE General test. A total of 10 different item types were administered within each form and parameter estimates were based on either the set of all verbal items (analogies, antonyms, sentence completions, and reading comprehension), all quantitative items (quantitative comparisons, data interpretation, and regular mathematics), or all analytical items (analysis of explanations, logical diagrams, and analytical reasoning). Each item was calibrated twice, once as an operational item and again when it appeared as a nonoperational item in the last section of a different test form than it initially appeared. The central focus of their study was to investigate the effect of exposure to the same item types earlier in the test on performance on certain item types (verbal, quantitative and analytical). They found that shifts in item positions seemed to affect difficulties of reading comprehension and data analysis items while practice effects canceled out with fatigue effects for the other item types. The effect of item shifts on item difficulty for analytical items was substantial. Further, Kingston and Dorans explored the differences in equatings as manifested by these items with different difficulty levels. Form B was equated to Form A twice using item response true score equating, once based on Form B parameters obtained when the items appeared in their operational locations, and again when the items appeared in the nonoperational location (the last section on either form) which mimics the precalibration equating design that requires resistance to item location effects. They observed that the equatings were different as analytical type-items were more sensitive to location changes. In their conclusion, they reiterated that for some item types, it is not recommendable to use precalibration data collection designs. “Likewise, these items types will be inappropriate for use in an adaptive testing context, using current item response models, because of their susceptibility to item location effects”(Kingston & Dorans, 1984, p. 153). On a final note, they suggest that whenever items exhibit within-test context effects, they should maintain the same location on the new form as they were in the old form.

In a related study to Kingston and Dorans’s (1984) research, Pommerich and Harris (2003) evaluated the effect of appended passage of pre-test items on reading and discrete pre-test mathematics items with respect to item statistics and examinee scores. For the reading passage, previously pre-tested items presented as a unit at the end of a test and had already been administered in an operational setting were later “re-field tested” using the same order of the passage unit as when the items were originally pre-tested and again using the same order as in operational test. However, there were minor changes made within the units. As for the mathematics items, a completely new pretest unit was administered using two different item orders. Using classical test p-values and item parameters from item response theory, the two researchers observed that items differed in difficulty, discrimination and guessing across the different administration conditions. They further concluded that the observed differences in performance might be due to changes in item position since they used randomly equivalent groups in this study and that the test administration condition were similar (controlled conditions). This research raises questions in terms of issues that may arise due to differences in pre-test administrations (appended versus embedded) such as examinee motivation in appended field testing and how much effect embedded field testing has on examinee operational score.

Despite having volumes of research on item position effects, the findings have not been consistent. Rubin and Mott (1984) investigated the effect of item position placement on item difficulty. Using a 60-item operational Reading test which comprised five subtests or competencies, experimental items were placed in three different positions (first, middle and last) on different test forms composed of like items. In all, they used sixty experimental items with each form having about 24 of the experimental items embedded in the test. Eighteen forms of the reading test were spiraled and administered to about 80,000 examinees. Rubin and Mott (1984) found out that the differences in mean difficulty values between the experimental items placed in the first place and the middle place, the first place and the last place, the middle place and the last place were .144, .049, and .095 respectively. In addition, the difficulty estimates of each item in one position compared to the difficulty estimates of the same item in a different position had a correlation of 0.95 or higher. Further, one-way analyses of variance indicated that there were no significant differences between mean Rasch item difficulties of the items placed in each of the three different positions. They concluded that the consistency of difficulty estimates of the items placed in different positions seems to support the notion that the position of an item in a test does not importantly affect Rasch item parameters. However, this study used Reading items which are for the most part, context dependent (Haladyna, 1992). This had an effect of moving clusters of items together which will not change the surrounding items much when the effect of movement of individual items are taken into account. With such limitations and conflicting findings with other researchers in the field, there is great need for renewed efforts in context effects in order to focus and exhaustively discuss important issues, like equating, that have much impact on testing today.

Most Recent Studies on Context Effects

Research on context effects has been evolving over the years. More so, along with state-of-the art innovations in testing such as item response models emerges the need to reassess and improve on issues about context effects especially on field-testing and operational testing. Among other areas affected with such changes in testing are item bank maintenance and test construction itself. Meyers, Kong, and McClarty (2008) evaluated the stability of item characteristics associated with items re-field tested to inform test policies pertaining item bank maintenance and test construction. They analyzed item characteristics for a set of test items in four subject areas and grade levels: Grade 4 Mathematics and Reading, Grade 7 Mathematics and Reading, Grade 8 Social Studies, and Grade 10 Science. These items had initially been field-tested in 2003 and were re-field tested in 2006. The purposes of their investigation were to determine how much item statistics had changed over a three-year period and to evaluate the practice of not including these three-year-old items on operational assessments. In built in this study was the fact that the items changed positions between different administrations. Little observed changes over time would suggest that the items should be eligible for inclusion on operational tests. On the other hand, significant changes will either suggest that such items should be retired from the item bank earlier or re-field tested to gather more updated item statistics prior to their inclusion on an operational assessment. For each grade and subject, the three investigators compared the two sets of item parameter estimates (2003 vs. 2006) by assessing the magnitude of differences in Rasch item difficulties and examining the correlations between the Rasch item difficulties, regressing changes in Rasch item difficulties on changes in item position and examining changes in item discrimination. They found out that the change in Rasch item difficulties did not change substantially when re-field tested in 2006 and that the correlations between the Rasch item difficulties across grades and subjects were highly positive, indicating a high correspondence between the two sets of indices. Also, and more in line with this current research, they found out that position change was not a statistically significant predictor of either item difficulty change or discrimination across grade levels and subject areas.

Also related to the preceding study, Davis and Ferdous (2005) investigated whether there are any differences in examinee performance on the same items administered to randomly equivalent groups during field-testing and later during live testing in Mathematics and Reading tests in Grades 3 and 5, with different test positions in the two administrations. They assumed that stability in item difficulty would suggest no effect of item location on examinee performance; i.e., an indication of the absence of test fatigue. Davis and Ferdous analyzed the data in two different ways. The first analysis aimed at finding overall mean differences in p-values and b-values between field and the live test items. They evaluated the stability of item difficulty estimates using analysis of variance. The second analysis investigated the difference in p-value and b-parameter between the field and the live test through Pearson correlation coefficients for the item difficulty estimates. The results from Analysis 1 (comparing changes in item difficulties by movement of items between blocks on the tests) showed that the relationship between item position and item difficulty on Grade 5 Reading only was statistically significant. However, in the second analysis, the correlation analyses for all tests except for Grade 3 Reading showed statistically significant results. The findings highlighted that there was a relationship between item repositioning on the tests and item difficulty. Davis and Ferdous (2005) further argued that the differences in conclusions between the two analyses might have occurred due to the limitations of the first analysis, which converted a continuous variable (item test position) into a binary variable (Block 1 or 2). In most situations, there is loss of information due to conversion of a continuous variable into a categorical variable resulting in statistical tests that may not find statistically significant differences.

Recently, He, Gao, and Ruan (2009) investigated whether pre-equating results agree with equating results based on operational data (post-equating) for a college placement program. They examined the degree to which IRT true-score pre-equating results agreed with those from IRT true-score post-equating and the results from observed-score equating. The three subjects that He et al. used in this study were Analyzing and Interpreting Literature (AIL), American Government (GOV), and College

Algebra (ALG). They found out that differences between equating results by IRT true-score pre-equating and post-equating varied from subject to subject. In general, IRT true-score post-equating agreed with IRT true-score pre-equating for most of the test forms. They concluded that any differences among the equating results were due to the way the items were pre-tested, contextual/order effects, or violations of IRT assumptions.

Several studies have investigated the effects of intact section rearrangements on test performance. Hill (2008) used classical test theory approach to investigate the effects of anchor item location on p-values by changing the order in which three reading passages appeared in a test. He found out that placing anchor items towards the beginning of a test results in higher p-values than when items appear towards the end of a test. However, the limitation of this study is that it involved moving around only three reading passages to change their locations. As a result, there were no large position shifts. However, the results highlight the worst-case scenarios that are more likely to manifest. The minimal shifts in position with large impacts in difficulty should be a source of concern whenever large position shifts occur.

In a bid to investigate the stability of the item-parameter invariance principle, Meyers et al. (2009) highlighted the proposition to weigh model advantages against concerns over threats to item parameter invariance when testing programs use IRT. They echoed Leary and Dorans’s (1985) point that item disclosure has affected the testing industry in that as the operational items are disclosed, field-testing items may not maintain the same positions when used as operational items. They further pin pointed the central problem that such changes in item positions do not only impact item difficulty but also affect equating results. In their research, Meyers et al. (2009) investigated the effect of item position on Rasch item difficulty (RID) in Mathematics and Reading using multiple regression where change in Rasch item difficulty was initially modeled as a function of item position change, grade, objective and time between field and live testing. However, upon inspection of the parameter estimates and t-test p-values, they dropped all independent variables thereby modeling Rasch item difficulty as a function of item position. Fitting cubic regression model to the data since it provided a better fit on both Mathematics and Reading revealed that 56% of the variance in change in Rasch item difficulty could be attributed to item position change in Mathematics whereas about 73% of the variance in change in Rasch item difficulty could be attributed to item position in Reading. Meyers et al. concluded that the regression model used indicated that placing an item nearer the end of the test has slightly more effect on its difficulty than placing it nearer the beginning of the test for Mathematics. As for Reading, the effect of placing an item nearer the end of the test has more severe effect on its difficulty than placing it nearer the beginning of the test. In general, average Rasch item difficulty changes decreased as item position decreased and increased as item position increased for both Mathematics and Reading. However, the limitation of this study is that it only analyzed data from untimed tests from one large K-12 testing program that utilizes a particular model, equating procedures, and test construction procedures.

In an attempt to address some of the limitations of the preceding study, Davey and Lee (2010) explored the existence of position effects in linear (non-adaptive) verbal and quantitative test forms assessing examinee’s readiness for graduate-level work. They also investigated whether pretesting items in random locations throughout the test may mitigate the presence of position effects (if they exist). Specifically, they investigated the effect of randomizing items in an unscored part of the test with 28 quantitative and 30 verbal items where examinees had no knowledge of the items that matter to their scores hence examinee motivation was not an issue. These items were sorted by difficulty (easy, medium, hard), item type (discrete items or passage based items) and by whether the items were consequential to examine the relationship between shift distance and difficulty change due to speededness. Using logistic regression and analysis of item residuals, Davey and Lee (2010) found that the relationship between shift distance and difficulty change is stronger for quantitative items than for verbal items and is evident when the easiest items appear towards the end of the test section. They also found that the relationship between shift distance and difficulty change is stronger for passage-based items than for discrete items and that test speededness is likely to affect the difficulty of quantitative items more than verbal items. For this study, position effects were noticeable only for the largest shifts in item position—those of half a test section length or more, Davey and Lee concluded that this validates the strategy of pretesting items in random positions.

Although Davey and Lee (2010) indicated that the more serious impact of item performance change is on the equating process, itself, Wise et al. (1989) present issues regarding potential interactions between the examinees' abilities and item positions are changed. More specifically, Wise et al. investigated the effects of item position on item statistics in a large data set for tests of word knowledge (WK) and arithmetic reasoning

(AR) using IRT parameter estimates and classical item statistics. Data were collected as part of a project to refine the Army's Computerized Adaptive Screening Test (CAST), an adaptively administered battery consisting of a word knowledge subtest and an arithmetic reasoning subtest. As part of this effort to refine the Army's Computerized Adaptive Screening Test, 275 new and existing items from the word knowledge and arithmetic reasoning subtests were administered to 20,071 Army recruits from five different Army posts. Two hundred seventy items for each subtest were divided into six non-overlapping sets of 45 items each and were then calibrated. The remaining five items were included in all six forms as potential anchors should subsequent equating prove necessary. Item statistics were computed separately for forward and reversed versions of each form. IRT and classical parameters were determined and the findings show that estimates for both parameters varied significantly with item position. However further investigation revealed that the variation was not generally related to the characteristics of the item, but was related to the ability of the examinees. Specifically, there were no significant position effects when average percent passing scores were 75% or higher while position effects were more evident when passing scores were 50% or lower. Although the process involved use of few linking items (5 out 45 items), which is even lower than the 20% threshold prescribed by Kolen and Brennan (2004) when common item equating is used, it does not change the conclusion that the item-parameter invariance principle, which is central to IRT, can be threatened when contexts change. In their conclusion, Wise et al. (1989) stressed on the need to embrace IRT methodology while being cautious and mindful of context effects.

Recently, Talento-Miller, Rudner, Han, and Guo (2012) examined data from an operational computer adaptive test program to determine the extent of possible position effects and differences by item type for verbal and quantitative sections with each section having two item types. They used data from a speeded operational test to determine if there were differences in item parameter estimation based on position of pretest item administration. They used pretest items that only appeared in both the beginning and ending sections; i.e., from the first or the last ten items. Talento-Miller et al. examined the differences in item statistics (*p*-values, median response time, and parameters based on Rasch and 3 PL models) based on the data from only the beginning and ending positions. Like many other researchers, they found out that items appeared more difficult when presented in end positions. In addition, the relative magnitude of differences varied by item type with response time. However, Talento-Miller et al. (2012) observe that the large effects observed in their study based only on the beginning and ending section will diminish based on item parameter estimates of the full set of responses from the beginning, middle, and ending sections.

In a related study, Meyers, Murphy, Goodman, and Turhan (2012) recently investigated the impact of item position change on item parameters and common item equating results under the 3 PL model. The study extends the Meyers et al. (2009) study by investigating the impact of item position change, sample size, subject area, grade, elapsed time between item uses and number of previous uses on changes in IRT a, b and c parameters using real and simulated data. In addition, Meyers et al., (2012) investigated the impact of item position change, sample size, subject area, grade, elapsed time between item uses and number of previous uses on the resultant D^2 statistic, defined as the weighted sum of the squared deviation between the item characteristic curves (Murphy, Little, Kirkpatrick, Fan, & Lin, 2010). Results of this study indicate that item position changes have a negative impact on item difficulty (b-parameters), discrimination (a-parameters) and D^2 statistics. In general, both a-parameters and b-parameters changed more the further an item shifted in position between administrations. Further, D^2 statistics had higher magnitudes the further items move from their field test locations. In addition, Meyers et al.'s (2012) simulation study has not only served to show that the D^2 values were very large in large shift conditions but has also illustrated how the derived scale scores differed more from their pre-equated values at each raw score point. In practice, such changes have the effect of increasing misclassification rates, an area that the current study intends to highlight. Further, the introduction of sample size and field test design in addition to the conditions studied using real data allowed full exploration and broader generalization for the impacts on the D^2 statistic. The incorporation of minor item position changes (placing items in same positions or within five positions from their field-test locations) and major item position changes (placing items within 9-15 positions from their field-test locations) in their simulation study mimic the best testing practice and worst case scenarios respectively, that can possibly be observed in real world testing situations.

Debeer and Janssen (2012) propose a new approach that combines DIF and linear logistic test models to detect and model the effects of item position and explore the use of IRT in descriptive and explanatory models (De Boeck & Wilson, 2004) for investigating and modeling the effects of item position on discrimination and difficulty. Traditionally, investigations of item parameter changes from different test forms/administrations take the form of analyses of test responses for different examinee groups followed by comparisons of item parameter estimates across groups. For example, in the Rasch context, many researchers (Meyers et al., 2009; Whitely & Dawis, 1976; Yen, 1980) have shown that items may differ in difficulty among test forms whose only difference is the position of the item in the test booklet. Debeer and Janssen (2012) consider such a finding as an instance of DIF for two groups of test takers defined by the test form taken by the different groups of test takers. Unlike the traditional two- step approach, the one-step approach that Debeer and Janssen propose has many benefits other than estimating all item difficulties simultaneously in one-step and placing them on the same scale. Among others and from a practitioner's standpoint, ability to model position effects allows for assessment of magnitude, shape, and direction of the effects. Moreover, the use of test characteristic curves (TCC) to derive test score for examinees allows model users to overcome any limitations that could arise from item-level modeling.

The model that Debeer and Janssen (2012) investigated assumes a linear "position effect" that augments the item difficulty. The model, derived from Fisher's linear logistic test model (LLTM), refined by Kubinger (2008, 2009), and later reformulated as the Random Weight Linear Logistic Test Model (RWLLTM) by Rijmen and De Boeck (2002) is as follows:

$$Logit\ [Y_{pik} = 1] = \alpha_i\,[\theta_p - (\beta_i + \gamma_p\ (k - 1))]$$

This is a two-parameter logistic (2PL) model where Y_{pik} is the probability of a correct response in each test form for person p, taking ith item in kth position. As the model shows, probability for a correct response is a function of test taker's ability, θ_p and item difficulty, β_{ik} as well as item discrimination α_i. In addition, γ is an additive increment to the item difficulty that reflects the item's position (representing a "drift" or "learning effect" when it is less than zero and a possible "fatigue effect" when it is greater than zero); γ_p is assumed to be a normally distributed random examinee effect with estimated mean $\mu(\gamma)$ and variance $\sigma^2(\gamma)$ (Debeer & Janssen, 2012). The addition of individual differences to the position effect (of item difficulty) makes the model practically useful since it allow the effect to be isolated for particular groups within the population. Accordingly, it may allow for the investigation of the effects of other person properties such as motivation and gender thereby allowing some neat interpretation that comes because of item placement. Overall, using their two illustrations, Debeer and Janssen (2012) found that test scores estimated for the model with an average position effect are lower than for the scores estimated for the model without a position effect. Further, for one standard deviation above the mean for position effect, the impact becomes larger while the impact of position effect for one standard deviation below the mean on test characteristic curve is that it shifts the TCC to be almost equal to the TCC of the model without position effects. At item level, Debeer and Janssen found that when γ_p is equal to the mean or one standard deviation above the mean, the position effect is positive and the success probability decreases. On the other hand, when γ_p is one standard deviation below the mean the position effect is negative. The implication of the two cases is that items become more difficult towards the end of the test when γ_p is equal to the mean or one standard deviation above the mean and become easier towards the end of the test when γ_p is equal to the mean or one standard deviation below the mean).

All in all, Meyers et al. (2012) and Debeer and Janssen (2012) examined non-speeded testing conditions. Large testing companies apply the commonly used rule of thumb—if 100 percent of examinees complete 75 percent of the test and 80 percent of the examinees complete 100 percent of the test, then the test is unspeeded (Debeer & Janssen, 2012). However, one can argue that test taker's knowledge of time limitations for speeded tests may exert time pressure that causes examinees to exhibit changes in their response patterns. Previous studies have indicated that examinees portray different behaviors at different points of the test. Such behaviors range from feeling no time pressure and spending more time at the beginning (Bergstrom, Gershon, & Lunz, 1994) to rushing at the end to catch up with time (Bridgeman & Cline, 2002; Talento-Miller & Guo, 2009). It is therefore worth investigating whether the present findings hold in speeded conditions.

Why Care about Context Effects?

In order to make sense of the amount of error associated with context effects, Haertel (2004) investigated the behavior of linking items in test equating through examining the magnitude of anchor-item selection effects on equating transformations. Using bootstrap and analytical procedures for estimating an error component representing the random selection of anchor items from a hypothetical pool of such items, he concluded that common item sampling constitutes a major overlooked source of error in equating. Wu (2010) extrapolated Haertel's findings by reporting that three sources of error are associated with estimated mean scores; error due to sampling of common items, error due to sampling of students and error in measuring individual students. Haertel computed the magnitudes of these sources of error relative to the percentage of the total error and found 83% of the error was due to sampling of the common items; only 11% was attributed to student sampling error, and 6% was accounted for as individual student error. The amount of common item sampling error variance is quite substantial. Similarly, Wu (2010) provides evidence, using IRT equating, that test booklet-order effects can have an effect size on average item difficulty of 0.4 or more.

The effects of item order in testing are even more striking when performance assessments (PAs) such as constructed response items, essays, oral presentations, exhibits, and portfolios are used. Although performance assessments are beyond the scope of this research, it is worth mentioning the impact that performance assessments have on testing. Muraki, Hombo, and Lee (2000) observe that although performance assessments are highly recommended for testing higher order thinking skills (e.g., synthesis and evaluation levels of Bloom's taxonomy) and are deemed to be more authentic than multiple choice items, they pose a lot of serious challenges for test equating and comparability of tests over time. Some of the challenges reported by Muraki et al. include context effects, practice effects, multidimensionality, small number of items resulting in inadequate sampling of construct domain, complexity of equating resulting from intra-judge rating inconsistencies and inter-judge rater severity differences, no useful anchor items and security problems due to the easy to memorize nature of items.

These challenges have resulted into increasing renewed efforts by researchers to investigate the applicability of existing methodologies to performance assessments and development of new methods for performance assessments. An educated glimpse into the future of testing may lead one to conclude that these channels of research in performance assessments will open new avenues of theories and methodologies that are more likely to revamp the whole testing industry.

Differential Item Functioning (DIF) and Testing

Ideally, a fair item is one that is comparably valid for all groups and individuals and that affords all examinees an equal opportunity to demonstrate the skills and knowledge that they have acquired and which are relevant to the test purpose (Roever, 2005). Test developers hope that examinees with the same ability on construct in question should perform similarly on test items despite having differences in other aspects such as gender, culture, ethnicity, religion and other factors. However, in practice, fairness at item and/or test level can become threatened with context effect issues (e.g., serial position item changes), resulting in item bias, which is characterized by test items having extraneous sources of difficulty that are not relevant to the construct being measured which impact test-takers' performance (Zumbo, 1999). Hambleton and Rodgers (1995) reiterate that the existence of bias is notable when a dimension on the test is deemed irrelevant to the construct in question, placing one group of examinees at a disadvantage in taking a test. Simply stated, construct irrelevant factors play a crucial role to an examinee's item responding behavior.

Traditionally, there have been many methods for detecting bias. Subkoviak, Mack, Ironson, and Craig (1984) mention that transformed item difficulty, the chi-square and the three parameter characteristic methods have previously been used to detect item bias. Currently, as Perrone (2006) indicates, the Mantel-Haenszel procedure; first developed for use in epidemiological research and later applied to the detection of differential item functioning (DIF) by Holland and Thayer (1988) is the standard of psychometric bias analysis. Usually, DIF analyses involve comparisons between the base (reference) group and the focal group where it is assumed that the two groups of examinees have the same ability for the construct that a given test is measuring. In such cases, it is easier to attribute any differences in examinee performance to differences arising from group composition. The existence of context effects due to serial position changes for multiple forms of a test implies that the response patterns of examinees who should otherwise be responding in a similar manner are affected. This has an impact on the quality of equating.

However, the presence of DIF is not always indicative of item bias. For example, when examinees have different abilities (e.g., lower ability and high ability examinees), it is expected that their responses on specific items will be different—differential item functioning is apparent. The difference in the performance of the two groups of examinees is due to disparate impact. The problem with detecting DIF for examinees with different ability levels is compounded by the fact that ability and DIF become confounded i.e., it is difficult to determine whether the difference in examinee performance is due to disparate impact or extraneous sources that are unrelated to the construct that is being measured. Wise et al. (1989) observed that there is differential impact between low and high ability examinees in different item arrangement schemes when speeded tests are used. While the occurrence of DIF due to disparate impact is a normal phenomenon in testing, test developers and users do not condone the existence of DIF due to extraneous sources as this affects validity of tests. The intolerance to DIF even worsens when extraneous sources differentially affect examinees.

As Zumbo (1999) elaborates, bias has considerable ramifications at policy, administrative and classroom level as it can lead to systematic errors that distort the inferences made in the classification and selection of examinees. Cook and Eignor (1991) observe that test scores are often used for such purposes as the assessment of the abilities and/or skills of individuals who are competing for college admissions or seeking professional certification. This evaluation of test scores (when used in conjunction with other information) may lead to a decision to exclude a candidate from some academic program or to limit the ability of an examinee to practice the profession of his/her choice. In addition, important funding decisions and other decisions regarding school curricula, etc., are sometimes dependent on the standardized test scores of groups of students. Therefore, with high stakes testing, it is important to pay attention to issues that may lead to bias (e.g., context effect issues) to ensure that different groups of examinees, irrespective of their ability levels, are provided with equal opportunity to excel and that classification accuracy levels of high standards are attained and maintained.

Summary

The adoption of an item response theory (IRT) model in practice requires accepting the underlying assumptions that the models. That said, psychometricians need to continually monitor and at least occasional evaluate the degree to which the model assumptions are empirically supported (Wise et al., 1989). This dissertation considers the degree to which context effects, item bank designs, test construction practices, and equating methodological choices challenge IRT equating assumptions.

As the reviewed literature suggests, context effect issues have been detected and are certainly very possible for any test design that allows items to migrate to new positions or slots in test forms over time. In this chapter, Leary and Dorans' (1985) comprehensive and historically documented review of the literature on context effects has spearheaded a variety of topics including but not limited to test equating that this study focuses on. Other researchers later echo many of the issues that Leary and Dorans (1985) raise about context effects.

The existence of context effects poses many technical equating challenges. As Kolen and Brennan (2004) put it, "clear cut criteria and rules for making equating decisions do not exist: The specific context of equating in the particular testing program dictates how these issues are handled. Equating involves compromises among various competing practical constraints . . ." (p. 268). Kolen and Brennan also point out that test design/development and equating are inseparable; however, the equating studies have often ignored possibly relevant test design and development issues. The present study attempts to fill some of that gap by directly addressing some test design issues that might interact with equating outcomes—for example, manipulating test design characteristics of the test form at large versus the characteristics of the anchor tests.

Some conflicting findings have been reported regarding the effects of item position shifts on examinee performance. Specifically, Rubin and Mott (1984) found that IRT Rasch item difficulty estimates did not seem to be impacted by item order. Conversely, Meyers et al. (2009) and Hill (2008) found that items moved towards the beginning of the test decreased in difficulty; the item difficulties also increased as the items were moved towards the end of the test. The conflicts have not been adequately resolved to date; however, studies like the current one may provide some new insights and may useful recommendations for test development practices related to equating.

CHAPTER III

DATA AND METHODOLOGY

This dissertation was comprised of two distinct phases. The first phase analyzed real data to explore the potential relationship between the type of item position changes (direction and magnitude of position shifts) and the extent of corresponding change (if any) in the item statistics. This real-data analysis provided the “reality check” on the second part of the study. The second part used a more elaborate set of IRT-based computer simulations to investigate the interaction of multiple test design conditions under manipulated item position effects. The impact of those manipulations was also considered when different types of IRT equating were performed.

Real Data

The first phase of this study involves an empirical analysis of data from two large-scale educational testing programs herein referred to as Assessment 1 and Assessment 2 programs, administered in third, fourth, fifth, sixth, seventh, eighth, tenth and eleventh grades. One thousand five hundred fifty seven mathematics items and one thousand six-hundred forty five reading items are analyzed to investigate the effects of serial position changes on item difficulty and examinees’ probability changes.

Although the two assessment programs have a common test developer and administer timed reading and mathematics tests in almost similar grade levels (see Figure 3.1), there were a number of differences between the two assessment programs. First, while both testing programs administer timed tests, assessment programs 2 appears to have more relaxed time demands to complete the tests than assessment program 1. For assessment program 2, with the exception of Grade 11 reading tests that should be completed in 60 minutes (and 30 additional minutes), the duration of all test sessions is 45 minutes (and 45 additional minutes) for all grade levels. On the other hand, the duration for assessment program 1 test sessions vary, the minimum duration is 25 minutes (no calculator math session) while the longest test session is 75 minutes (session B math). Reading test durations are between 40–60 minutes for assessment programs 1. However, additional 20 minutes are allowed for the mathematics and reading tests for assessment program 1.

The second difference between the two assessment programs is that the number of items and item types per session are different for similar grade levels. For example, assessment program 2 has three reading sessions, each with 14 multiple-choice (MC) items and three constructed-response (CR) items for third and fourth grade. Assessment program 1 has three reading sessions for similar grades as assessment program 2 but the number of items per session is different (with minimum of 18 and a maximum of 30 MC items).

Finally, there are three cut points along each score scale, established via content-based standard setting procedures, that are used to categorize every examinee into one of four proficiency categories: (a) warning, (b) partially proficient examinees, (c) proficient, and (d) advanced. The two assessment programs obviously have different scales for reading and math and different cut points.

Although this study focuses only on the MC items, the presence of factors such as item types, differences in session length and timing could affect examinee performance. The data were segregated by program to reflect some of those design and standardization differences. In total, 580 mathematics items were analyzed from assessment program 1 and 977 items were analyzed from program 2. Similarly, there were 702 reading items in program 1 and 943 items in program 2. The tests were administered over five consecutive academic years. Table 3.1 shows a summary of the grade levels involved in this study and the total number of items used per subject and grade.

Table 3.1

Number of Items Used and Grade Levels Where the Assessments Were Administered from 2005-2010 Academic Years

Assessment Program	Grades in Mathematics	Number of Mathematics Items	Grades in Reading	Number of Reading items
1	3,4,5,6,7,8 and 11	580	3,4,5,6,7,8 and 10	702
2	3,4,5,6,7,8 and 11	977	3,4,5,6,7,8 and 11	943

An embedded field item design helps ensure that examinees are unaware of the items' status (operational versus field test). Accordingly, lack of motivation did not appear to be a likely cause. These assessments are calibrated and equated using the IRT three-parameter logistic (3PL) model. The test developers employ classical item analyses and IRT calibration criteria to identify problematic field test items, possibly suggesting answer key changes, or that the items be eliminated from operational use. Content and statistical specifications are employed to assemble the test forms, however, subject-matter experts approve the content and organization of the final test forms, and psychometric experts verify that the proposed tests have the desired statistical characteristics.

Analyses Procedures Using Real Data

It was important to calculate the magnitude of item positions changes for the examinations (e.g., direction and positional distances moved) between the initial use of the items and subsequent reuse during the five consecutive years. The item's initial administration counted as the "base year." For any subsequent use of an item that was used previously, the base administration position was subtracted from the new position to get an item position change value (Δ_p). A positive change indicates that the item moved towards the end of the test; a negative position change indicates that the item moved towards the beginning of the test. A visual inspection of the data shows that a majority of the items did not change positions across forms and or administrations over the five academic year period. However, there were some significant position changes with some items moving over 40 position slots. That magnitude of position change is equivalent to moving an item from one end of the test to the other. The minimum replicated uses of an item across test forms was two (field test and then operational). The maximum use count across forms/ administrations for some items was seven.

Because the assessment programs use 3PL response model, item statistics (difficulty, discrimination, and guessing parameters) were obtained following calibrations that linked these statistics to the bank scale. The real data analysis explores the extent to which the difficulty measures (i.e., b-parameter estimates) changed from one form/administration to another when position changes were either changed or remained the same. Like position changes, this process involved computing a simple difference between pairs of b-parameter estimates for the same item (e.g., the difference between the b-parameter estimated during field testing and the b-parameter estimated during operation use of an item). A negative difference in the item difficulties indicates that an item was easier on the subsequent administration(s) than for the initial administration. A positive b-value difference indicates that the item appears to have gotten more difficult during the subsequent administration(s) than for the initial administration.

Additionally, using the 3 PL model,

$$Prob(u_i = 1|\theta) \equiv P_i(\theta) = c_i + (1 - c_i)\{1 + exp[-a_i(\theta - b_i)]\}^{-1},$$

the probability of a random examinee with θ equal to one of the three cut points delineating the four proficiency categories established for this examination (warning, needs improvement, proficient, and advanced) was computed. The differences in the conditional probabilities, $P_{i(\text{second})}(\theta_k)–P_{i(\text{base})}(\theta_k)$ were computed by subtracting from the "second" use the "base" probability. A positive difference in the conditional probabilities indicates that the item appeared to get easier for examinees having proficiency scores in the region of that cut score, θ_k (k = 1, 2, or 3). A negative difference in the conditional probabilities indicates that item got more difficult for examinees having proficiency scores in the neighborhood of the cut score. Given corresponding changes in the 3PL item discrimination parameter estimates (a_i), it is theoretically possible for the same item to simultaneously exhibit positive and negative changes in the conditional probabilities at different cut scores if the estimated item characteristic curves cross.

In this study, conditional probability differences were only computed relative to the “base” use of each item. For example, items that were administered seven times—that is, administered on seven different forms/administrations throughout the five-year testing period—had six conditional probability differences computed at each of the three cut scores (18 differences in total) all computed relative to the initial item administration. The differences in the conditional probabilities at each of the three cut scores were subsequently compared relative to the magnitude of changes in the item position (if any). Table 3.2 shows the template that summarizes the computation process for the probability differences at the three cut points.

Table 3.2

Template for Calculating Probability Differences at the Three Cut Points

Item ID	Position Dif (Δ_p)	Cut 1	Cut 2	Cut 3
1	0	…	…	…
1	0	…	…	…
1	4	…	…	…
2	0	…	…	…
2	15	…	…	…
2	-7	…	…	…
3	0	…	…	…
3	-2	…	…	…
4	-40	…	…	…
4	-36	…	…	…
4	10	…	…	…
4	20	…	…	…

From the template, one can deduce that even if an item maintained or changed position on different occasions (form/ administration), probability differences were still computed for the different occasions the item was administered. The advantage of using three distinct cut scores is that it provides a mechanism for evaluating conditional (versus marginal) effects of context (e.g., serial item position changes) on examinee responding behavior at different proficiency levels. The probability differences (i.e., differences in the response functions) also avoid concerns over the choice of IRT model and potential changes in all of the model parameters, since any of those interactions are reflected in the estimated IRT response probabilities.

To illustrate the differences that occur at cut points—potentially because of context effects (e.g., serial position changes)—the exploratory part of the study also investigates the differences among different proficiency levels for different position changes. Essentially, this involved creating item bins based on observed item movement distances. Somewhat consistent with "item movement distance" categories proposed by Davey and Lee (2010), seven item position bins were formed for the mathematics data from the second assessment program. Three of the bins were used for items that moved *forward* towards the beginning of the test; i.e., 16 positions and above, between eight and fifteen, and between one and seven positions. Similar categorizations were used for items that moved *backwards* towards the end of the test i.e., 16 positions and above, between eight and fifteen, and between one and seven. The seventh category comprised items that did not change positions. Item position would typically be treated as an ordinal variable; putting the items into discrete bins formed more stable ordered categories variable (i.e., with larger item frequencies per bin). However, forming the bins—that is, increase the bandwidth of intervals along the item position scale) could also cover up some legitimate variation.

In the present case, some item position bins had as few as 70 items re-administered for some movement distances. In order to somewhat overcome the instability of the statistics estimated for the smaller-frequency bins, bootstrap sampling was employed, stratified by bin, to stabilize the estimated statistics and provide a closer picture to population dynamics.

Because the real data did not afford the opportunity to know "truth", majority of the phase 1 analyses are correlational-based in nature—that is, statistically associating the apparent magnitude and direction of change in item difficulty (or differences in the conditional probabilities) with observed changes in the magnitude and direction of item positions.

Simulated Data

The second phase of the study uses computer-generated data simulations to specifically manipulate the nature and extent of item difficulty and then investigate: (a) conditions that lead to worst/best testing situations, (b) the effect of item difficulty change on equating, and (c) the impact of changes in item difficulty on score accuracy and decision/classification accuracy. Although the amount and nature of change in difficulty introduced in the simulation study is derived from the real data analyses, it is worth mentioning here, that as it is typical with other simulation studies, the complexity of real data is not fully reflected in the current simulation study and may therefore lead to indefinite generalizations and conclusions. However, apart from allowing many experimental conditions to be set up, controlled, and replicated, which may not be feasible with real data, the simulation study allows discernment of patterns and trends that could otherwise be difficult to capture with real data.

Item Bank Generation and Calibration

Two data sets were created by randomly sampling examinees proficiency scores from a normal ability distribution, $\theta \sim (\mu, \sigma^2)$ with $N = 5000$ in each case. The item responses were created using GEN3PL equating version of response generation software (Luecht, 2012b). One data set was created for a base form (BF) and the second data set for an alternate form (AF) using separate, user-specified item parameters for each form; that is, (i) there is a file containing the BF vectors of item parameters, $\{\mathbf{a}_{\text{BF}}, \mathbf{b}_{\text{BF}}, \mathbf{c}_{\text{BF}}\}$ and (ii) another file containing the AF vectors of item parameters $\{\mathbf{a}_{\text{AF}}, \mathbf{b}_{\text{AF}}, \mathbf{c}_{\text{AF}}\}$. A subset of items in both the BF and AF was designated as a common-item anchor set (CIAS). Depending on specific study conditions considered, the CIAS item parameters (especially the b-parameters) either had the same values in both the BF and AF files, or different values for the AF values to stimulate difficulty changes and other contexts or “difficulty effects” as a function of serial position or other factors (refer to table 3.3). The IRT-based data simulations employed the three-parameter logistic (3PL) item model;

$$Prob(u_i = 1|\theta) \equiv P_i(\theta) = c_i + (1 - c_i)\{1 + exp[-a_i(\theta - b_i)]\}^{-1}.$$

For these simulations, manipulations for the distributions of item discrimination and difficulty parameters were necessary to alter the measurement precision of the test design in prescribed ways relative to the distribution of examinee scores and/or key decision points along the proficiency scale while maintaining an average guessing parameter of about 0.15 for all tests.

Description of Simulation Conditions

Table 3.3 provides the conditions included in the simulations. There are two types of conditions. Some of the conditions relate to test design (i.e., location and amount of IRT measurement information) and affect the precision of scores or decision accuracy. Other conditions directly involve the manipulation of item drift for the anchor test items. From Table 3.3, average item discrimination, item difficulty, and test length relate to the whole test while proportions of anchor items, magnitude of drift and correlation between re-administered and original difficulty relate to anchor items.

Table 3.3

Simulation Conditions

Conditions	Detail	Count
Average discrimination: whole test	mean(*a*)=.6, mean(*a*)=1.0	2
Item difficulty: whole test	*b*~(0,1), *b*~(-1,1), *b*~(0,.6)	3
Test length: whole test	*n* = (50, 100)	2
Proportions: anchor test length	*p* = (0.2, 0.3)	2
Magnitude of drift: anchor test	*d* = (-.25, 0.0, .25)	3
Correlation: re-administered and orig. difficulty	*r* = (0.8, 1.0)	2
Total conditions		144

In general, average item discrimination is directly related to the precision of test scores. Lower discrimination values imply lower precision and larger measurement errors while higher discrimination values imply higher precision with smaller measurement errors. Two levels of item discrimination were considered here, a = .6 and a = 1.0. Operationally, this study refers these levels of item discrimination as moderate (denoted mod_a), and high (denoted high_a), respectively.

In addition, in IRT, the item difficulty distributions also relate to the location of the measurement precision (potentially highest near the mean of the item difficulty distribution) and the spread of the precision (with a larger standard deviation spreading out the information over a larger range of the proficiency scale). This study investigated three levels of mean difficulty and variability in difficulty. The first level denoted as *b ~ (0, 1)* indicates that measurement precision is targeted at a mean of zero and allowed to spread (vary) with 1 standard deviation. In this study, this level is referred to as normally distributed difficulty or sometimes as moderate difficulty with reasonable variability. The second level denoted as *b ~ (-1, 1)*, shows that measurement precision is targeted at a low level, mean of -1 and allowed to have variability of 1 standard deviation just like the previous level. Operationally, this level is termed low mean difficulty with reasonable variability. The final level for difficulty and variability is targeted at a mean of zero and has variability of 0.6 standard deviations. This level is denoted as *b ~ (0, .6)* and for purposes of clarity in this study is described as moderate mean difficulty with constricted variability.

Likewise, test length relates to measurement precision, where longer tests tend to be more precise than shorter tests. The proportion of the test that is comprised of anchor items is important as a larger proportion determines the extent of influence the anchor test has on the equating, as well as the stability for the equating results. In this study, four test designs are investigated. The first test design has a test length of 50 and includes 10 anchor items. This test design is denoted 50_10 test in this study. The second test design, like the first test design also has 50 items. However, the number of anchor items is different from the first design as it has 15 anchor items. This test design is denoted 50_15 test. The third test design, denoted 100_20 has 100 items in all, 20 of which are anchor items. Finally, the fourth test design consists of 100 items from which 30 items are anchor items. Consistent with the current nomenclature, this test design is denoted 100_30 design. In line with prior research, the four test designs meet the minimum 20 percent anchor item requirements as either 20 percent or 30 percent representation was initially maintained for the test designs.

Further, varying the magnitude of difficulty change for anchor items provides a direct way to introduce difficulty change impacts into the generating model hence mimicking the impacts of context effects such as position effects discussed earlier. In this study, b-values for the CIAS on the base form are manipulated by decreasing their mean difficulty by .25, increasing the mean difficulty by .25, or leaving the CIAS to have the same mean difficulty. This results in three alternate forms with different mean anchor difficulty conditions, herein referred to as negative change in mean anchor item difficult (*b_delt < 0)*, no change in mean anchor item difficulty, (*b_delt = 0)* positive change in mean anchor item difficult (*b_delt > 0).*

Finally, beyond adding or subtracting a constant to the difficulty parameters to mimic drift, the magnitude of correlation between original b-values and b-values after re-administration is also indicative of the amount of agreement between the two item administrations. For this study, two levels of correlations are manipulated. The first level of correlations is referred to as moderate correlation, denoted mod_r, which depicts situations where the relationship between CIAS on base form and alternate form is not very strong and correlation is set between 0.75 and 0.8. On the other hand, a correlation of 1 implies that there is a strong relationship between CIAS on base form and alternate form. This is also known as high correlation and is denoted high_r.

All examinee scores in these simulations are sampled from a unit-normal distribution—that is, $\theta \sim N\ (0, 1)$. Although additional sampling distributions are interesting to explore, the current number of conditions in the simulation study limits expanding those conditions to include characteristics of the examinee population(s). In essence, using a unit-normal distribution for all examinees provides a very nice baseline where all examinees are randomly sampled from the same population—implying that the groups truly are randomly equivalent and any equating should statistically maintain a reasonable similarity among the equated score distributions.

Despite the use of the 3PL model for data generation, all of the item calibrations and equating steps were carried out using a Rasch IRT =(i.e., a one-parameter logistic) model:

$$P_i(\theta) = [1 + exp(b_i - \theta)]^{-1}.$$

Using the 3PL for data generation and the Rasch model for the calibrations and equating steps incorporates a reasonable amount of model-misfit into the simulations—that is, statistical *noise* that would be plausibly observed with real data.

Further, the base form response-data are locally calibrated in WinSteps (Linacre & Wright, 2011) to get the bank scaled item difficulty values that are used for equating purposes. This is the same type of WinSteps analysis (unanchored, unconstrained) regardless of the equating method or any other study conditions. These estimated base form item difficulties from WinSteps and not the "true" item difficulties from the generating model are used to represent the item characteristics relative to a common bank scale (θ). On the other hand, the alternate form response-data are used to conduct the equatings that are fully discussed in the following sections.

One of the minor complications of Rasch model IRT calibrations is that the most popular Rasch calibration software, WinSteps centers the item difficulties at zero for a *local* (i.e., unconstrained) calibration. The reason for this centering is to facilitate calibration of an item bank via the item difficulties, rather than examinee scores. However, for forms that differ in difficulty, this choice of centering method implies that the averages of the proficiency score estimates would be different, even if the examinee samples taking each test form were randomly equivalent. Equating to the bank scale should theoretically adjust for the centering of any particular test form. Therefore, in this study sampling examinees from a unit normal distribution creates a scaling problem with the easy and reasonable difficulty-variability condition, b~(-1,1).

In order to linearly adjust the true thetas to put them on the same scale as the WinSteps-estimated thetas, the following transformation was used:

$$\theta_{true*} = \beta_0 + A\ (\theta_{true}),$$

where θ_{true} is the generated ability parameter from GEN 3PL, *A* is the slope parameter and is obtained by computing the reciprocal of mean discrimination, i.e., *A=1/mean (a_i)* for the test form of interest. If *mean (a_i)* =1, then *A=1*. This implies that, as *mean (a_i)* <1.0, the adjusted true theta scores will spread out. Alternatively, as *mean (a_i)* >1.0, the distribution of adjusted true theta scores will decrease in spread. β_0 is the intercept and is computed as follows:

$$\beta_0 = mean\ (d_i) - A^*\ mean\ (b_i),$$

where *A* is as described above, d_i is the Rasch difficulty with *mean (d_i)* i.e., local calibration, *mean (b_i)* is the average difficulty for the generated parameters for the test form of interest. Finally, the adjusted abilities (θ_{true*}) are used for computing the residuals for this difficulty and variability level.

Rasch Equating Methods

Rasch equating methods are prevalent among test developers. This study investigates equating results under the simulation conditions outlined in the prior sections for three linking/equating methods. Specifically, the study compares and contrasts precision accuracy in retaining original ability estimates when *unweighted mean equating* (UME), *weighted mean equating* (WME) *and anchor item calibration* (AIC) linking/equating methods are used. Furthermore, the UME and WME methods are implemented using one of two ways to treat the CIAS. First, using the full CIAS—i.e., using the complete unaltered set of common items with each of the three methods of equating, denoted CIAS.F in table 3.4. Second, a subset consisting of only the most statistically stable items in the CIAS is used, denoted CIAS.S in table 3.4, where the stabilization process is carried out using a statistical stability analysis described in later sections. As a result of this process, the UME and WME methods of equating are combined with the two treatments of the CIAS to form a two-by-two set of fully crossed equating/linking method conditions as shown in Table 3.4. Combined acronyms are shown in the individual cells (e.g., UME-CIAS.S implies doing unweighted mean equating with a statistically stabilized CIAS). In all, the fully crossed two-by-two set of equating method formed by UME and WME plus the single treatment anchor-item calibration result in five equating methods.

Table 3.4

Factor Breakdown to Produce Five Equating Method Conditions

	TREATMENT OF CIAS	
EQUATING METHOD	Full CIAS Used CIAS (F)	Stabilized CIAS Used CIAS (S)
Unweighted Mean Equating (UME)	UME-CIAS.F	UME-CIAS.S
Weighted Mean Equating (WME)	WME-CIAS.F	WME-CIAS.S
Anchored Item Calibration (AIC)	AIC-CIAS.F	__________

Ideally, the UME process involves the computation of an equating constant based on the average differences between the base form and alternate form items in the common item anchor set. The BF item difficulties are assumed to exist from a prior calibration. The equating constant is then added to all of the items on the alternate form. A final anchored calibration is then run to get ability estimates (maximum likelihood estimates or MLEs of θ. The equating constant, Δ, is computed as follows;

$$\Delta = \mu\left(\widehat{b_B}\right) - \mu\left(\widehat{b_L}\right),$$

where B = bank scale and L= local calibration scale. All anchor test items are included when computing the mean. The equating constant is then added to all of the difficulty estimates for the local calibrations to put all of the items on a common scale[1], that is;

$$b*_{j,AF} = b_{j,\,AF} + \Delta$$

where $b_j*_{,AF}$ denotes the new item difficulty after adding a constant, Δ, to the item difficulty, $b_{j\,AF}$ from the common item anchor set that is *locally* calibrated in WinSteps for all $j = 1 \ldots n$ items on the alternate form. Once the UME constant, $\Delta_{\mathrm{UME},}$ is computed and applied to all of the alternate form items, the statistically adjusted (equated) b_j* values are fixed in a final WinSteps calibration of the alternate form response data to obtain MLEs on the base form. Finally, the MLEs are compared to the “true” generated θ values (generated using GEN3PL equating version) in the calibrated sample.

[1] When external (unscored) anchor test items are used, an additional recentering step is needed that centers only the scored items, prior to computing the mean equating constant.

Like the UME process, the WME process assumes that the BF item difficulties exist from a prior calibration and finds an additive constant, $\Delta_{\text{WME,}}$ which is added to locally calibrated alternate form item difficulties to put them on the bank scale. However, unlike UME, in WME, the contribution towards the mean from each item is determined by the standard error associated with the item. The equating constant is computed as;

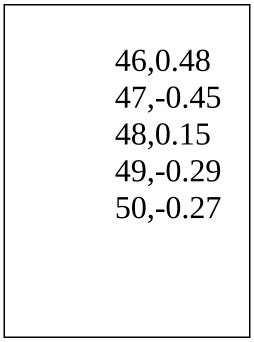
$$\Delta_{WME}= \sum_{i\in CIAS}^{n_{CIAS}} W_{i,BF} b_{i,BF} - \sum_{i\in CIAS}^{n_{CIAS}} W_{i,AF} b_{i,AF}$$

where, $b_{i,BF}$ denotes an item difficulty from the CIAS previously calibrated to the bank scale and $b_{i,AF}$ is the item difficulty from the same CIAS items locally calibrated in WinSteps (i.e., item difficulties for all of the items in the AF form are centered at zero as is the case with the UME method without anchoring the calibration). The item weights, $W_{i,\ BF}$ and $W_{i,\ AF}$ are statistically optimal (normalized) weight functions that Graybill and Deal (1959) demonstrated would produce composite mean, in this case, the means of the BF and AF CIAS items with minimum variance properties. Stated differently, anchor item difficulty estimates are weighted by their reciprocal error variance estimates to provide an unbiased equating constant with minimum error variance properties (Graybill & Deal, 1959). The weights are computed using the error variances of estimate:

$$W_{i,f} = \frac{S_{i,f}^{-2}}{\sum_{i\in CIAS}^{n_{CIAS}} S_{i,f}^{-2}},$$

where, $S_{i,f}^{-2}$ is the reciprocal error variance of estimate for the CIAS items with f = (BF or AF). This approach effectively weights the equating constant more heavily for those items having more stable estimates of b_i. Some states have adopted this approach for computing the mean equating constant as an alternative to mean equating. Cohen, Jiang, and Yu (2009) highlight that the volatility of mean equating becomes clear when the standard error of the item difficulty parameters is considered—few students get very difficult items correct, and likewise, few students get easy items wrong. Ideally, weighted linking operates on the principle that the weight of imprecisely measured items on the final linking constant is reduced i.e., reducing the weight associated with items with large standard errors.

Finally, anchor item calibration (AIC) involves directly linking a set of new items to a bank scale (or to another test form) by fixing the CIAS item difficulty estimates in the alternate form at their base form estimates. WinSteps then estimates item difficulties (*b*-statistics) for only the unanchored items. In WinSteps, AIC is carried out by: (1) generating a CIAS file that contains an item index (position) and b-statistic for each of the CIAS items and (2) adding the IAFILE= CIAS_filename command to the WinSteps control file, where CIAS_filename is the anchor item file name created earlier. For example, if we created a CIAS file with five anchor items in the last positions of a 50-item test, it might look like Figure 3.1, where for instance, -0.45 is the calibrated item difficulty from base form, here appearing in position 47.



```
46,0.48
47,-0.45
48,0.15
49,-0.29
50,-0.27
```

Figure 3.1. A sample CIAS file for facilitating anchor item calibrations in WinSteps.

Stability Analysis

Stability analyses were carried out using 1PLIWgtEqt Version 1.0 software (Luecht, 2012). This software conducts Rasch *stability analysis* by carrying out an iterative sequence of unweighted mean equating steps that estimate an equating constant, Δ, such that;

$$\Delta = \frac{\sum_{i=1}^{n_{anchors}} b_{i(bank.scale)} - \sum_{i=1}^{n_{anchors}} b_{i(local.calib)}}{n_{anchors}}$$

where “bank.scale” refers to the base-form calibrated or item-bank scale and “local.calib” refers to the new test form(s) locally calibrated (e.g., calibrated in WinSteps without item anchoring so that the mean of all the items is centered at zero). The equating constant is used to update the locally calibrated item difficulties,

$$b^{*}_{i(local.calib)} = b_{i(local.calib)} + \Delta$$

The second step involves computation of a displacement statistic, δ, which is a measure of item difficulty difference between the base form and alternate form and is expressed as follows:

$$\delta_i = \left| b_{i(bank.scale)} - b^*_{i(local.calib)} \right|.$$

At each iteration, the item with the maximum value of δ is eliminated from item linking (anchor) set until *max*(δ) > τ, where τ is a user-defined "maximum acceptable displacement." By convention in the Rasch literature, τ usually equals .3 or .5 (Luecht, 2012a). Note that if *max*(δ) ≤ τ at the first, iteration, all of the linking (anchor) items are retained. The mean equating constant, Δ, after the first iteration is also the usual mean equating constant (no items eliminated). To be consistent with the literature, the value of τ was set at 0.3 in this study.

Further, as an output file, 1PLIWgtEqt generates an anchor item file, *UnWeighted-WS-Anchor file* after the final iteration corresponding to WinSteps required item_position, b_value format. As a final step to complete the analysis, the anchor item file is directly specified in the WinSteps control file using the *IAFILE= anchor. file* specification in order to link the two tests using only the stable items.

In a similar style, 1PLIWgtEqt also conducts a stability analysis using a weighted-means equating, where the weighted mean for the "bank.scale" and "local.calib" item difficulties are weighted by an "information weight" (Graybill & Deal, 1959). That is,

$$\Delta_W = \frac{\sum_{i=1}^{n_{anchors}} w_{i(bank.scale)} b_{i(bank.scale)} - \sum_{i=1}^{n_{anchors}} w_{i(local.calib)} b_{i(local.calib)}}{\sum_{i=1}^{n_{anchors}} w_{i(bank.scale)} + \sum_{i=1}^{n_{anchors}} w_{i(local.calib)}}$$

where

$$w_i = \frac{1}{s^2_{\hat{b}|b}},$$

with the error variance of the estimated b_i denoted as

$$s^2_{\hat{b}|b} = \left[SE\left(b_i\right)\right]^2.$$

SE^2 denotes the squared standard error of estimate from the calibration. Just like UME stability analysis, WME stability analysis involves applying the constant Δ_w to all locally calibrated items and compute the displacement statistic, δ. At each iteration, the item with maximum value of δ is eliminated from item linking (anchor) set until all the remaining items in the link set have a displacement statistic less than 0.3. The program generates an anchor item file, *WeightedMn-WS-Anchor.file* after the final iteration corresponding to WinSteps required item_position, b_value format. As a final step to complete the analysis, the anchor item file is directly specified in the WinSteps control file using the *IAFILE= anchor. file* specification in order to link the two tests using only the stable items.

In summary, 1PLIWgtEqt generates three output files. The first output is a detailed output file showing all iteration steps for the stability analysis and results for both the unweighted and weighted analyses showing the items that have been pruned from the common item anchor set, which are still part of the test as unique items. The second output, *UnWeighted-WS-Anchor file*, is the final (post-stability analysis) item anchor file from the unweighted mean stability analysis whereas the third output file is the *WeightedMn-WS-Anchor file* which is the post stability analysis from the weighted mean stability analysis. Both post-stability analysis files are then specified in WinSteps using the IAFILE control command to execute an anchored item calibration using only the post-stability analysis items to complete the analysis.

Evaluation of Equating Accuracy

There are a number of ways for evaluating item parameter changes. For the real data, the primary focus of parameter change is with respect to conditional probability values (i.e., changes in the probability of a correct response), $\Delta_p(\theta) = P_1(\theta) - P_0(\theta)$, where the subscripts respectively denote the initial (0) and next (1) use of an anchor item. Serial position is the magnitude of change in the item sequence: $\Delta_{i\,=}\, i_1 - i_0$. For the simulated data, residual statistics play an important role. Known proficiency scores (generated proficiency scores), θ_j, for j=1… N simulated examinees (where θ_j is the ability of the jth person and N = 5000) are compared to the estimated proficiency scores (maximum likelihood estimates, $\hat{\theta}_J^{ML}$), computed using the Rasch model using three evaluation criteria.

First, the average residual, $\overline{\delta} = \hat{\theta}_J^{ML} - \theta_j$for the 5000 examinees is analyzed to provide an index of bias, which is expected to be zero. The average bias is expressed as follows:

$$\overline{\delta} = \frac{1}{N}\sum_{j=1}^{N}(\widehat{\theta_J^{ML}} - \theta_j),$$

where, as already stated, $\hat{\theta}_J^{ML}$ is the estimated ability for examinee j obtained from WinSteps and θ_j is the true ability for the same examinee (or alternatively another examinee with similar ability) generated using GEN 3PL program.

Since the expectation of bias is zero, the second evaluation criteria, the root mean squared error (RMSE), provides a measure of variability of the differences in generated and estimated ability scores and is expressed as follows:

$$\sqrt{\frac{1}{N}\sum_{j=1}^{N}(\hat{\theta}_J^{ML} - \theta_j)^2}$$

Finally, the known proficiency scores and ML estimates are used to determine true versus estimated classifications with respect to fixed decision points along the score scale. Specifically, misclassification percentages for two cut scores are examined mimicking two cut points when real data is used; not proficient/proficient cut point, set at θ = -0.5 and proficient/advanced cut point which is set at θ = 1.0. These cut scores are obtained from real data situations. The generated proficiency scores are used as the true classifications and the estimated ability scores for the simulated conditions are compared to the generated proficiency scores resulting in percentage classification rates.

CHAPTER IV

RESULTS

Chapter Overview

The primary purpose of this study is to examine the impacts of item parameter changes for anchor test items on Rasch equating results and the quality of score estimates for performance-related decisions. Using real and simulated data, four research questions guided this study. The analyses of the real data explored the magnitude and extent of item parameter changes by addressing the first research question: How does the magnitude and direction of item difficulty changes and conditional probability changes relate to serial item position changes at different proficiency levels on the ability scale?

In addition, item statistics from real data analyses provided plausible values of the sampling distribution of item statistics for the simulated-data study. Accordingly, the organization of this chapter has the results presented in that same sequence. Importantly, the simulated data, because “truth” is known and errors or residuals can be directly computed for every simulated examinee and every item, provide additional information where the influence of the manipulations of the simulation conditions and/or equating methods can be directly linked to functions of those residual errors. In review, three research questions guided the simulated data analyses. First, the simulated data section investigated the comparability of three IRT Rasch model equating methods and two treatments of the anchor item sets (fixed and iteratively stabilized). Further, the simulated data results section discusses whether there is any gain in using stability-based equating over fixing all of the anchor items during the equating process.

In addition to the residual-based outcomes, simulated data results are also provided regarding decision classification accuracy (and inaccuracy) rates. The use of multiple cut scores on the “true” IRT θ scale make these results relevant to certain types of mastery tests as well as educational testing applications that report examinee proficiency outcomes in terms of achievement levels (e.g., basic, proficient and advanced).

Real Data Results

Relationship between Item Position Changes and Difficulty Changes

Table 4.1 shows the correlations between change in item positions and changes in difficulty values when re-administered items maintained or changed their positions for both mathematics and reading in the two assessment programs. In general, there was a tendency of items becoming easier when item positions changed from a later position in the test to an earlier slot. Conversely, items became harder when moved from an early position to a later slot. Clearly, such findings are more notable in reading than in mathematics items for the two assessment programs. The results are inconclusive for mathematics however, where moving an item towards the beginning of the test had the effect of making it easier in one assessment program while the same phenomenon had the effect of making items harder in the other program. For assessment program one and two in Reading, about 5% and 7% of the variance in changes in item positions accounted for changes in item difficulty respectively. For mathematics, only about 1% of the variance in changes in item positions accounted for changes in item difficulty for both assessment programs.

Table 4.1

Correlation between Change in b Parameters and Change in Position

Assessment Program	Subject	Correlation (sig)	Item Re-admin
One	Mathematics	.070 (.018)	1145
	Reading	.228 (.000)	1942
Two	Mathematics	-.096 (.000)	2398
	Reading	.268 (.000)	1457

Correlation significance level in parenthesis

Table 4.2 shows difficulty change descriptive statistics for the two assessment programs for selected item position changes with higher count of re-administered items in mathematics. Items that did not change positions exhibited the smallest mean changes and had little variability in difficulty from one administration to another for the two assessment programs. With no change in item positions, the effect size was 0.13 and .06 for assessment programs one and two respectively. However, as observed before from the correlation tables, items appeared to be more difficult when placed towards the end of the test and easier when placed towards the beginning of the test for assessment program one. In contrast, items appeared to be easier when placed towards the end of the test and more difficult when placed towards the beginning of the test for assessment program two.

Table 4.2

Descriptive Statistics for Difficulty Changes for the Two Assessment Programs in Mathematics

Program	Δ Position	Re-admin Count	*M*	Min	Max	*SD*
	-13	10	.028	-.233	.295	.172
	-7	9	-.054	-.258	.256	.215
	-5	60	.134	-.351	.928	.227
	-3	75	-.017	-.426	.623	.192
One	0	474	.005	-.172	.333	.039
	3	57	.109	-.622	2.01	.445
	5	60	-.055	-1.025	.549	.292
	10	7	.110	-.044	.265	.122
	17	15	.269	-.225	.788	.294
	-19	21	-.011	-.549	.284	.171
	-8	21	.106	-.393	.902	.336
	-5	55	-.040	-.776	.262	.215
	-2	161	.017	-.881	.481	.182
Two	0	658	-.010	-1.437	.809	.170
	2	150	.014	-.671	.803	.244
	5	43	.040	-1.007	.671	.288
	17	18	-.189	-.705	.256	.241
	23	55	-.148	-.963	.578	.316

Figure 4.1 illustrates the relationship between mean difficulty changes and position changes for all items re-administered in different positions for the two

assessment programs. As was the case with the selected cases discussed earlier, with the exception that the mean changes and variability are minimal when no position changes occur, the pattern in mathematics seems to be inconclusive.

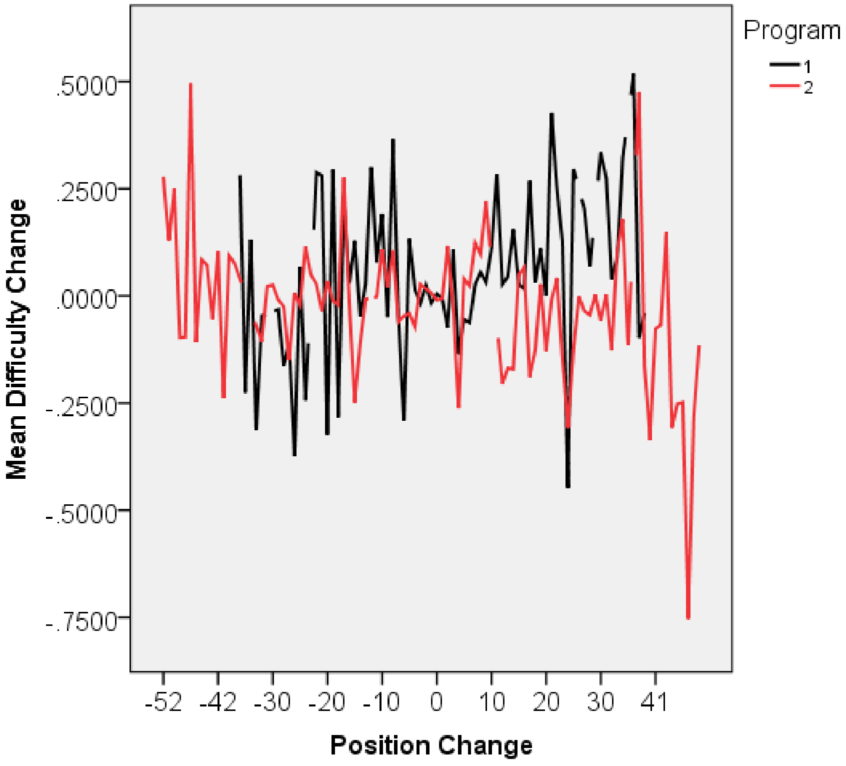


Figure 4.1. Relationship between mean difficulty changes and position changes for all items re-administered in different positions for the two assessment programs in Mathematics.

Table 4.3 shows difficulty change descriptive statistics for the two assessment programs for selected item position changes with higher count of re-administered items in reading. As was the case with mathematics, items that did not change positions exhibited the smallest mean changes and had little variability in difficulty from one administration to another for the two assessment programs. With no change in item positions, the effect size was 0.03 and 0.05 for assessment programs one and two respectively. However, as observed before from the correlation tables, items appeared to be more difficult when placed towards the end of the test and easier when placed towards the beginning of the test for the two assessment programs.

Table 4.3

Descriptive Statistics for Difficulty Changes for the Two Assessment Programs in Reading

Program	Δ Position	Re-admin Count	*M*	Min	Max	*SD*
	-17	12	-.124	-.394	.073	.207
	-15	29	-.188	-1.026	.306	.282
	-7	142	.106	-.575	.512	.199
	-3	15	-.022	-.564	1.202	.574
One	0	903	-.003	-1.526	.517	.106
	5	11	-.062	-.019	.176	.055
	7	120	-.057	-.549	.766	.236
	10	33	-.023	-.157	.229	.122
	15	36	-.236	-.142	.669	.261
	-34	39	-.168	-.777	.216	.093
	-17	106	-.034	-.892	.156	.094
	-10	8	-.300	-.458	.253	.113
	-1	73	-.069	-1.150	.327	.116
Two	0	722	.011	-.717	1.296	.202
	2	19	-.088	-.434	.325	.183
	17	91	.037	-.871	.714	.301
	18	19	.168	-.200	.506	.201
	34	37	.150	-.222	.821	.201

Figure 4.2 further illustrates this pattern using all re-administered items. Reading items were harder if located towards the end of the test and were easier if located towards the beginning of the test. Unlike in mathematics where the findings were inconclusive, the general pattern in reading was clear.

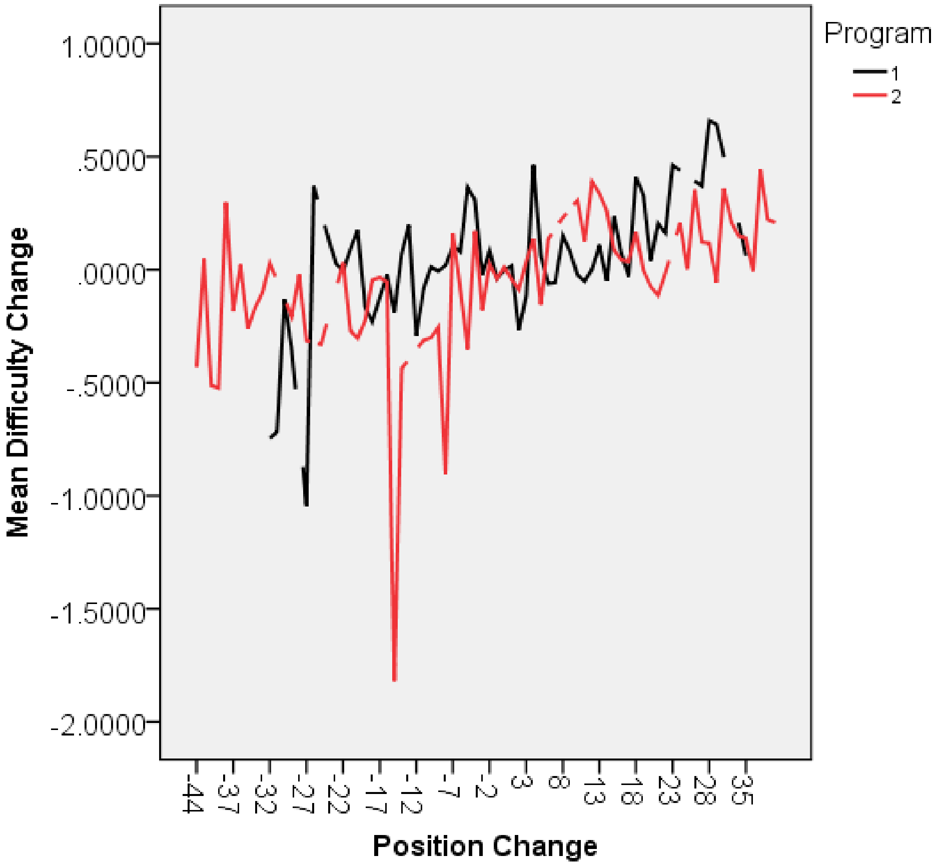


Figure 4.2. Relationship between mean difficulty changes and position changes for all items re-administered in different positions for the two assessment programs in Reading.

In addition to item difficulty analyses for mathematics and reading, similar analyses for probability changes for examinees at the three cut points were conducted and the results are shown in Table 4.4. In general, there is a significant relationship between changes in position and changes in examinee’s probability to respond to an item correctly. Specifically, when items are taken earlier in the test, examinees are more likely to respond correctly compared to the situation where items are administered towards the end of the test which results in lower probability for a correct response. However, this relationship seems to be more elaborate in reading than in mathematics and at lower proficiency levels. The lower the cut point the stronger is the negative relationship between examinee’s probability of responding to an item correctly and item position changes. As proficiency level gets higher, the negative relationship between position changes and changes in probability becomes less pronounced. This indicates that lower ability examinees are highly affected by changes in item positions than higher ability examinees.

Table 4.4

Correlation between Changes in Probabilities and Change in Positions at the Three Cut Points

Assessment Program	Subject	Correlation (sig)			Item Re_admin
		Cut point 1	Cut point 2	Cut point 3	
One	Mathematics	-.100 (.001)	-.073 (.013)	-.061 (0.039)	1145
	Reading	-.086 (0.00)	-.074 (.001)	-.013 (.565)	1942
Two	Mathematics	-.040 (0.05)	-.010 (.620)	.067 (.001)	2398
	Reading	-.145 (.000)	-.077 (.003)	.060 (.023)	1457

Note. Correlation significance level in parenthesis

Table 4.5 shows probability change descriptive statistics for the two assessment programs for selected item position changes with higher count of re-administered items in mathematics. Items that did not change positions exhibited the smallest mean change and had little variability in probability change from one administration to another for the two assessment programs. With no change in item positions, the effect size was 0.03 and 0.01 for assessment programs one and two respectively. In general, the effect size is larger for large position changes than it is for smaller position changes. Similar to the findings from correlations between item position changes and probability changes, the mean probability change seems to be lower for items that appeared towards the end of a test and is higher for items that occupy earlier slots.

Table 4.5

Descriptive Statistics for Probability Changes for the Two Assessment Programs in Mathematics

Program	Δ Position	Re-admin Count	*M*	Min	Max	*SD*
	-13	10	-.008	-.292	.200	.156
	-7	9	.040	-.077	.226	.113
	-5	60	-.033	-.288	.224	.129
	-3	75	-.026	-.178	.262	.091
One	0	474	-.003	-.319	.306	.095
	3	57	-.022	-.166	.183	.089
	5	60	-.013	-.241	.170	.103
	10	7	-.124	-.232	.034	.098
	17	15	-.099	-.437	.127	.156

Table 4.5 (cont.)

Program	Δ Position	Re-admin Count	*M*	Min	Max	*SD*
	-19	21	-.016	-.151	.216	.093
	-8	21	-.020	-.164	.156	.094
	-5	55	.011	-.275	.253	.113
	-2	161	.013	-.424	.327	.116
Two	0	658	-.001	-.392	.455	.105
	2	150	.009	-.381	.339	.125
	5	43	-.001	-.366	.399	.132
	17	18	.006	-.158	.189	.102
	23	55	.007	-.246	.339	.131

As Figure 4.3 shows, the findings in mathematics are not so definitive for the two assessment programs. However, this figure clearly illustrates that items that moved ten positions towards the beginning of the test and ten positions towards the end of the test (from -10 to +10 position changes) show little or no change in probabilities whereas outside this position change range (-10 to +10), there is more variability in probability changes.

Table 4.6 shows probability change descriptive statistics for the two assessment programs for selected item position changes with higher counts of re-administered items in reading. As has been discussed earlier, items that did not change positions exhibited the smallest mean change in probability change from one administration to another for the two assessment programs. With no change in item positions, the effect size for change in probabilities was merely 0.01 and 0.02 for assessment programs one and two respectively. Similar to the findings in mathematics, the mean probability changes in reading seem to be lower for items that appeared towards the end of a test and is higher for items that occupy earlier slots.

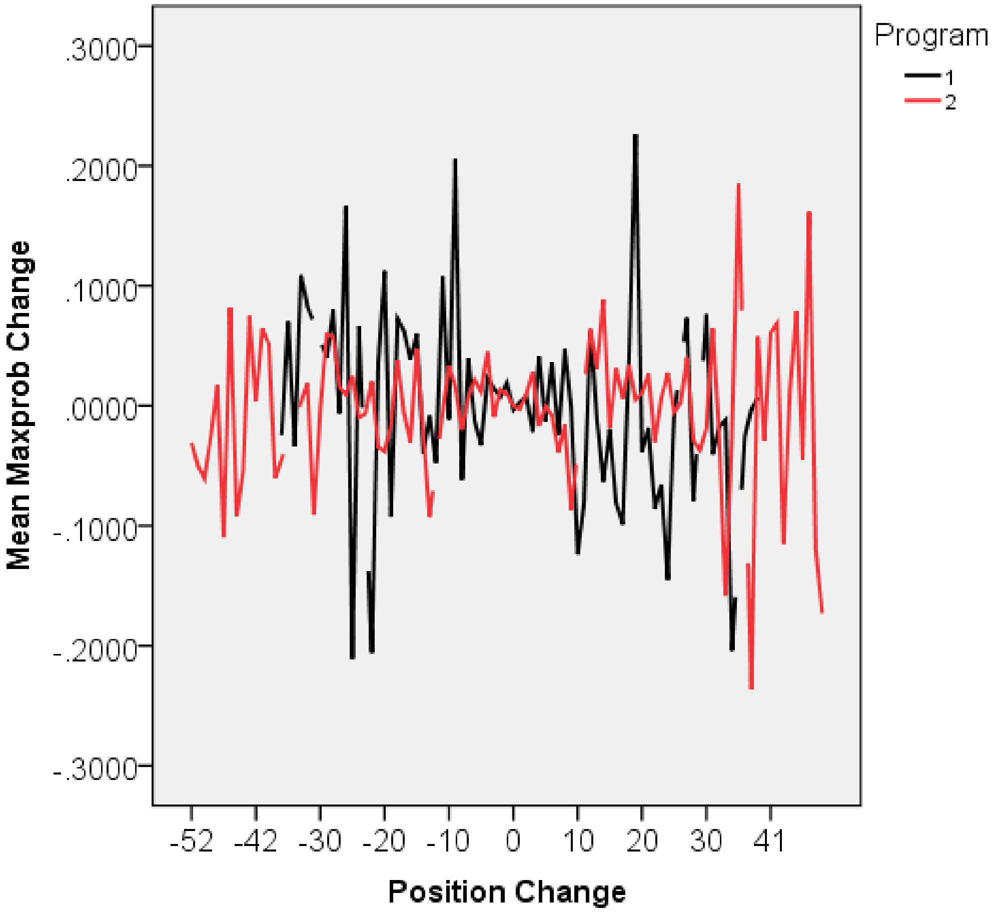


Figure 4.3. Relationship between mean probability changes and position changes for all mathematics items re-administered in different positions for the two assessment programs.

With reading items, the pattern seems more defined as Figure 4.4 indicates that for larger negative position changes the probability changes were higher while for larger positive position changes the probability changes were lower. Stated in other words, and consistent with the foregoing discussion, examinees' probability of responding to a given item correctly increased as the item was moved to an earlier position in the test and it decreased as the item was moved towards the end of a test form.

Table 4.6

Descriptive Statistics for Probability Changes for the Two Assessment Programs in Reading

Program	Δ Position	Re-admin Count	*M*	Min	Max	*SD*
	-17	12	-.043	-.472	.110	.176
	-15	29	.023	-.258	.270	.123
	-7	142	-.012	-.438	.460	.168
	-3	15	.009	-.139	.135	.103
One	0	903	-.002	-.504	.528	.134
	5	11	.034	-.141	.296	.115
	7	120	.009	-.479	.406	.170
	10	33	.052	-.241	.398	.158
	15	36	-.071	-.437	.252	.165
	-34	39	.042	-.153	.275	.091
	-17	106	.002	-.298	.334	.113
	-10	8	.115	-.003	.231	.084
	-1	73	.022	-.387	.426	.106
Two	0	722	-.002	-.413	.498	.111
	2	19	.005	-.351	.230	.138
	17	91	.003	-.385	.285	.109
	18	19	-.024	-.338	.130	.103
	34	37	.021	-.208	.175	.091

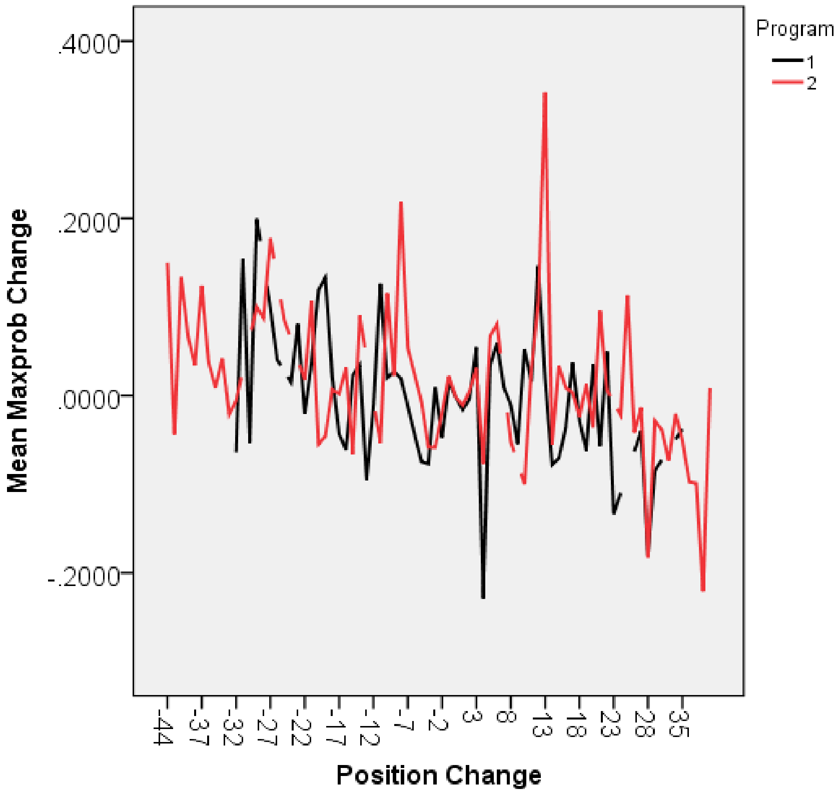


Figure 4.4. Relationship between mean probability changes and position changes for all reading items re-administered in different positions for the two assessment programs.

Figure 4.5 shows box and whisker plots for probability differences for low ability examinees from cut point 1 (blue) and high ability examinees from cut point 3 (green) for different position movement-distances for mathematics items. In general, administering items in the same positions in subsequent tests shows that examinees' probabilities of responding to the items have the least variability and is about zero for both low and high ability examinees. However, although the magnitudes of variability are low and the probability differences about zero, the two groups exhibit some differences at no position change. The larger variability in probability changes for low ability examinees shows that some examinees still find the items harder even though there is no change in position.

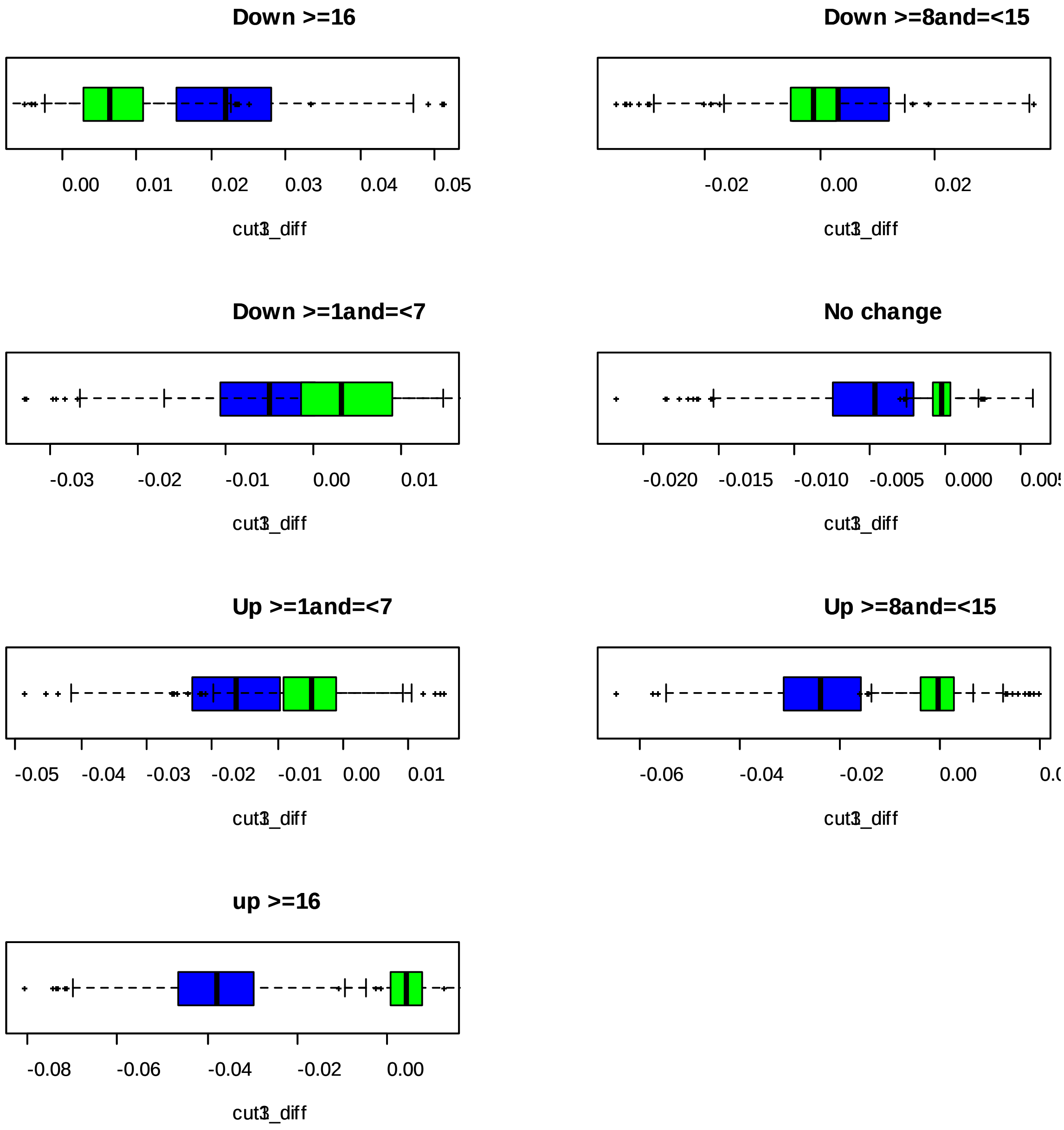


Figure 4.5. Probability differences between high ability examinees and low ability examinees for different item movement distances.

On the other hand, the low variability in probability changes for high ability examinees indicates that for this group of examinees, items are equally challenging when they occupy same spots during re-administration. Because there is no change in item position, the observed differences may be because of test wiseness between the two groups of examinees. Therefore, if item position changes do not play a role in the way examinees respond to items, a similar pattern should manifest for all position movement-distances that occur in both directions of the test—towards the end or the beginning of a test.

However, different position movement-distances do not have similar pattern to that observed for no position change. The effect of moving items towards the end of the test conspicuously manifests for lower examinees probabilities. The probability differences for low ability examinees decrease as the items move further up the test signifying that items got harder for this group of examinees. However, there is not much of an impact for higher ability examinees as items move further up the test. Although there is a lot of variability for high ability examinees when items occupy later slots compared to the no change situation, the probability differences are about zero for all position movement distances. As Figure 4.5 illustrates, the two groups of examinees become very distinct from each other with large position movement-distances (indicated by the amount of spaces between the box plots). Overall, the patterns for position movement-distances towards the end of the test are not similar to that of no change case. Because the patterns are not similar to the baseline case (no change situation), position differences seem to have an effect on examinees probabilities as position movement-distances increase towards the end of the test.

On the other hand, as position movement-distances increase towards the beginning of a test, probability differences for low and high ability examinees are positive i.e., items generally became easier for the two groups of examinees especially when position movement-distances were larger than eight. As was the case with position movement-distances towards the end of a test, the two groups of examinees exhibited differences in effects due to probability differences for position movement-distances towards the beginning of the a test. However, the probability differences for low ability examinees are more positive than for high ability examinees for items that moved eight positions or further. This implies that putting items towards the beginning of the test increased the probability of low ability examinees to respond to the items correctly more than it did for high ability examinees. While the effect of position changes manifest differently for the two groups of examinees, the separation of the two groups is not as distinct as when position movements were towards the end of the test.

Overall, Figure 4.5 shows that when items change positions, they become either difficult or easier depending on the direction and magnitude of the change. Apparently, these changes in difficulty become very notable for low ability examinees in comparison to high ability examinees. Because high ability examinees are already more likely to get most items right, it is more unlikely to notice any changes due to changes in difficulty and /or context effects. To the contrary, with low ability examinees, there is a lot of room to investigate the impact the difficulty of an item has on an examinee; many low ability examinees are already missing many items and therefore decreasing or increasing the difficulty of an item enormously affects the probability of these examinees to respond to the item correctly.

Simulated Data Results

Bias and Root Mean Squared Error

As discussed previously, bias statistic is used to measure the extent to which the ability estimates align with those of the generating model after equating. The rationale behind this comparison is that if the ability estimates from the calibrations and equating methods are consistently showing no difference to true scores, the study assumes an inconclusive premise due to study conditions. On the other hand, any differences between generated and estimated abilities will spell that the study conditions are having an impact. Just to reconfirm, the study acknowledges that the generating model is different from the model used for estimation. However, the presence of such discrepancies mimics the noise that is always present in real world testing situations. The study also acknowledges that although rescaling of b~ (-1,1) conditions was conducted to reflect the characteristics of the original population from which data was sampled (through the process discussed earlier), it could not be done without some amount of measurement error.

The second research question seeks to investigate the comparability of different study conditions and Rasch equating methods in terms of adequacy to attaining successful equating within and across test designs. Further, the third research question seeks to verify whether there is an observable difference between stabilizing anchor items or leaving them as fixed. Again, the effects are examined within and across test designs for similar study conditions. To effectively discuss these two research questions, similar study conditions within test designs are discussed in terms of bias and root mean squared error (RMSE) followed by an examination of similar conditions across test designs.

Table 12 in Appendix A shows the proportions of anchor items to the total number of test items after stabilization. It is clear that conditions with moderate correlation between anchor item difficulties have more items pruned thereby making the proportion of anchor items to fall below the recommended limit of 20% in many cases.

Bias and RMSE for 50_10 Test Design

The 50_10 test design consists of 50 items in all, ten of which are anchor items. Figures 17-22 in Appendix C show modified box plots that depict the mean, the first quartile, the third quartile, and the two extreme bias values on both sides of the distribution (minimum and maximum values). Clearly, there is about zero bias for all test conditions where mean item difficulty for the whole test is moderate with reasonable or constricted variability i.e., b ~ (0, 1) or b ~ (0, 0.6). However, there is a slight tendency of overestimating the Rasch ability estimates for all equating methods especially for conditions where change in mean anchor difficulty is negative (b_delt= -). In addition, when study conditions are as stated above, there is a decreasing trend in bias when change in mean difficulty for anchor items on alternate forms changes from negative to positive (from b_delt < 0 to b_delt > 0). For instance, for study conditions with moderate discrimination (mod_a), moderate correlation (mod_r) between anchor item difficulty on base and alternate forms, and where mean item difficulty for the whole test is moderate with reasonable, b ~ (0, 1), bias for *b_delt > 0* condition is lower than for *b_delt < 0* condition (see Figure 17). Similarly, the range of bias becomes less for *b_delt* condition. This also applies to conditions with moderate mean difficulty and reasonable variability test conditions.

However, the largest amounts of bias are observed for all conditions where mean item difficulty for the whole test is low with reasonable variability, b ~ (-1, 1). Under these conditions, the mean bias ranges from about -1 (when less discriminating items are used) to almost +1 (when high discriminating items are used). For the most part, the estimated Rasch ability values are underestimated and the middle 50 percent bias values have more variability when less discriminating items are used under this difficulty-variability level. On the other hand, when high discriminating items are used the ability values are generally overestimated and the middle 50 percent bias values have less variability. Just like the b~ (0,1) and b~(0, 0.6) conditions, bias trends decrease when change in mean anchor-item difficulty on alternate forms increases for b ~ (-1, 1) condition.

Further, within study conditions, the three equating methods with two treatments on UME and WME seem to be performing in the same way in terms of bias. Even when worst and best study conditions are considered (as discussed earlier), the amount of bias for the three equating methods with two treatments still remains similar within study conditions. To this effect, there is no noticeable precision gain for using anchor item stabilization over fixed number of anchor items. More specifically, the amount of mean bias when either UME-fixed vs. UME-stabilized or WME-fixed vs. WME-stabilized are used is similar. Therefore, for this test design, conclusions that uphold the use of one equating method over the other, or stabilized equating over fixed equating cannot be drawn.

Figure 4.6a shows the amount of root mean squared error (RMSE) for various study conditions by equating method for the 50_10 test design when item discrimination remains constant i.e., moderate (mod_a). RMSE is between 0.5 and 0.75 for all conditions with moderate mean difficulty and reasonable or constricted difficulty variability, b ~ (0, 1) or b ~ (0, 0.6). However, for similar conditions that differ only when mean item difficulty for the whole test is low and the variability is reasonable, b ~ (-1, 1), RMSE is between 1 and 1.25. Strikingly though, when all conditions are held constant and mean anchor-difficulty change (b_delt) is manipulated, RMSE decreases as mean anchor difficult change is becomes positive, i.e., from negative b_delt to positive b_delt. Therefore, RMSE is highest for negative b_delt and lowest for positive b_delt when all other study conditions remain constant. Within study conditions, the effect of using different equating methods is not noticeable. Thus, no precision gain in using stabilized equating over fixed equating methods.

On the other hand, all conditions with high discriminating items (high_a), exhibit similar patterns as those shown by condition with low discriminating items discussed in the preceding paragraph. As Figure 4.6b indicates, RMSE is between 0.5 and 1 for all conditions with moderate mean difficulty and reasonable or constricted difficulty variability, b ~ (0, 1) or b ~ (0, 0.6). However, high discriminating conditions with low mean item difficulty for the whole test and reasonable variability, b ~ (-1, 1) exhibit high uniform RMSE which values of about 1.

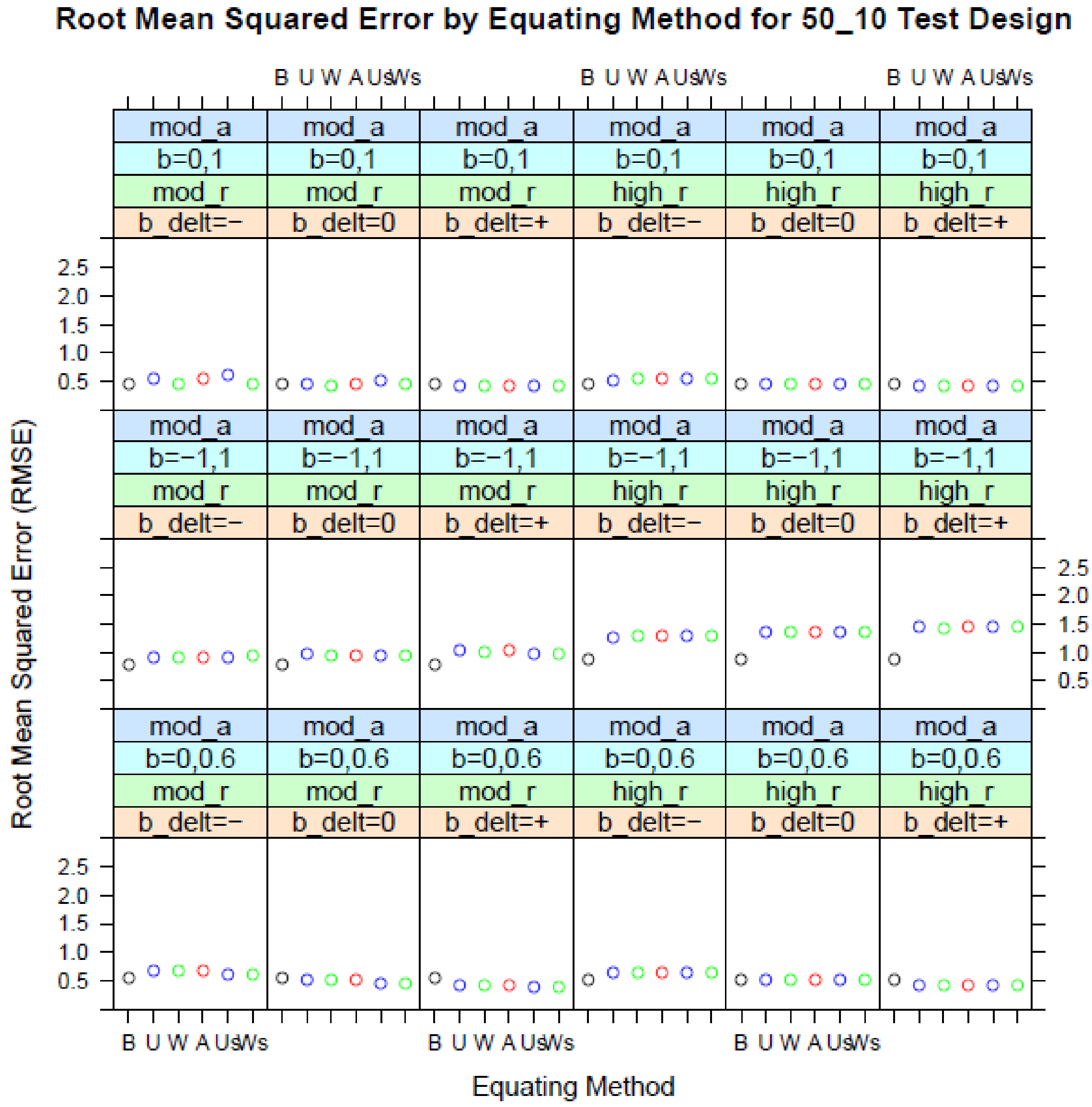


Figure 4.6a. Root mean squared error for moderate item discriminating conditions by equating method for 50_10 test design. (B = Base form, U = unweighted mean equating, W = weighted mean equating, A = Anchor Item Calibration, U_s = stabilized unweighted mean equating with stabilization and W_s = stabilized weighted mean equating with stabilization)

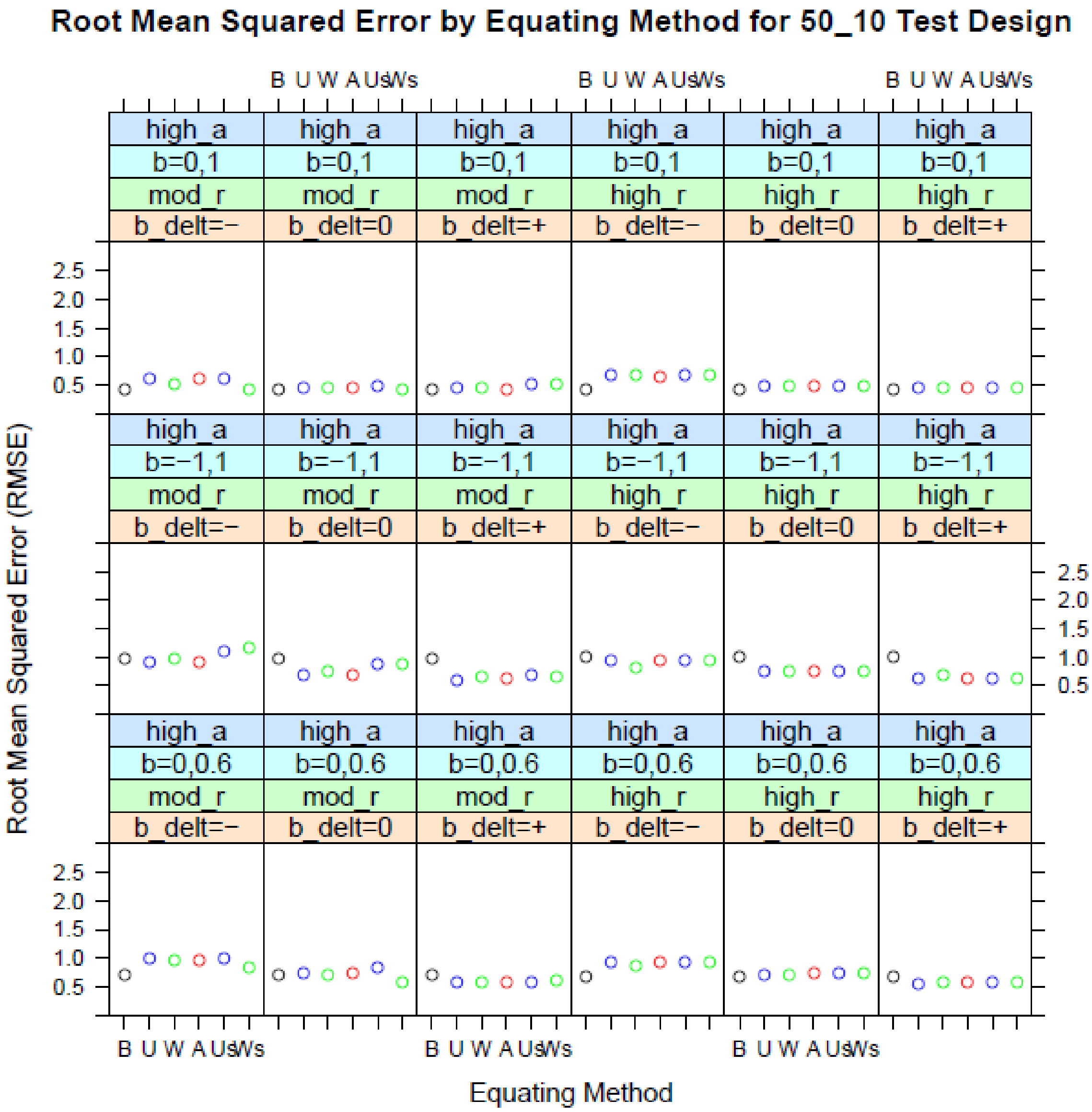


Figure 4.6b. Root mean squared error for high item discriminating conditions by equating method for 50_10 test design. (B = Base form, U = unweighted mean equating, W = weighted mean equating, A = Anchor Item Calibration, U_s = stabilized unweighted mean equating with stabilization and W_s = stabilized weighted mean equating with stabilization)

The main difference between the two discriminating levels is that RMSE variability for moderate item discriminating conditions are higher than their high discriminating item-conditions counterparts especially when b ~ (-1, 1). Other than this, all other effects that apply to moderate item discriminating conditions apply for high item discriminating conditions. Specifically RMSE is highest for negative b_delt and lowest for positive b_delt when all other study conditions remain constant, except when b ~ (-1, 1), for high discriminating conditions. Within study conditions, the effect of using different equating methods is not prevalent. Therefore, there is no precision gain in using stabilized equating over fixed equating methods.

Bias and RMSE for 50_15 Test Design

The 50_15 test design consists of 15 anchor items and 35 unique items. From the modified box plots in Figures 23–28 in Appendix C, there is about zero bias for all test conditions where mean item difficulty for the whole test is moderate with reasonable or constricted variability, b ~ (0, 1) or b ~ (0, 0.6). However, there is a slight tendency of overestimating the Rasch ability estimates for all equating methods especially for conditions where change in mean anchor difficulty is negative (b_delt= -). Also, as was observed in the 50_10 test design, there is a decreasing trend in bias when change in mean difficulty for anchor items on alternate forms changes from negative to positive (from b_delt < 0 to b_delt > 0). Therefore, when all other study conditions remain constant and b_delt varies, the amount of bias is always at the lowest for positive b_delt and highest for negative b_delt for b ~ (0, 1) and b ~ (0, 0.6) study conditions.

On the other hand, larger amounts of bias are observed for all conditions where mean item difficulty for the whole test is low with reasonable variability, b ~ (-1, 1), high discrimination and *b_delt* is negative. Under these conditions, the mean bias ranges from about 0.8 and 1. Less discriminating conditions coupled with low mean item difficulty for the whole test and reasonable variability, b ~ (-1, 1) generally register negative bias values on average but with more variability for the middle 50 percent bias values (see the length of the boxes in Figure 24 in Appendix C). This shows that the values of ability estimates are underestimated using the Rasch equating methods.

Further, within study conditions, the five equating methods generally seem to be performing in the same way in terms of bias. However, although there is no clear pattern to conclude that one equating method is more effective than the other methods, or to conclude that anchor item stabilization is better/worse than fixed item equating, differences in amounts of bias for the equating methods within study conditions are driven mainly by the level of correlation of anchor-items. Within study conditions, all equating methods are susceptible to higher differences in mean bias when moderately correlated anchor item conditions prevail. Nevertheless, the differences in bias for the equating methods within these study conditions do not follow a discernible pattern that can lead to conclude that one equating method is more effective than the other methods.

Figure 4.7a shows the amount of root mean squared error (RMSE) for various study conditions by equating method for the 50_15 test design when moderately discriminating item conditions are used. RMSE is between 0.4 and 0.9 for all conditions with moderate mean difficulty and reasonable or constricted difficulty variability, b ~ (0,

1) or b ~ (0, 0.6). However, for similar conditions that differ only when mean item difficulty for the whole test is low and the variability is reasonable, b ~ (-1, 1), RMSE is slightly above 1.

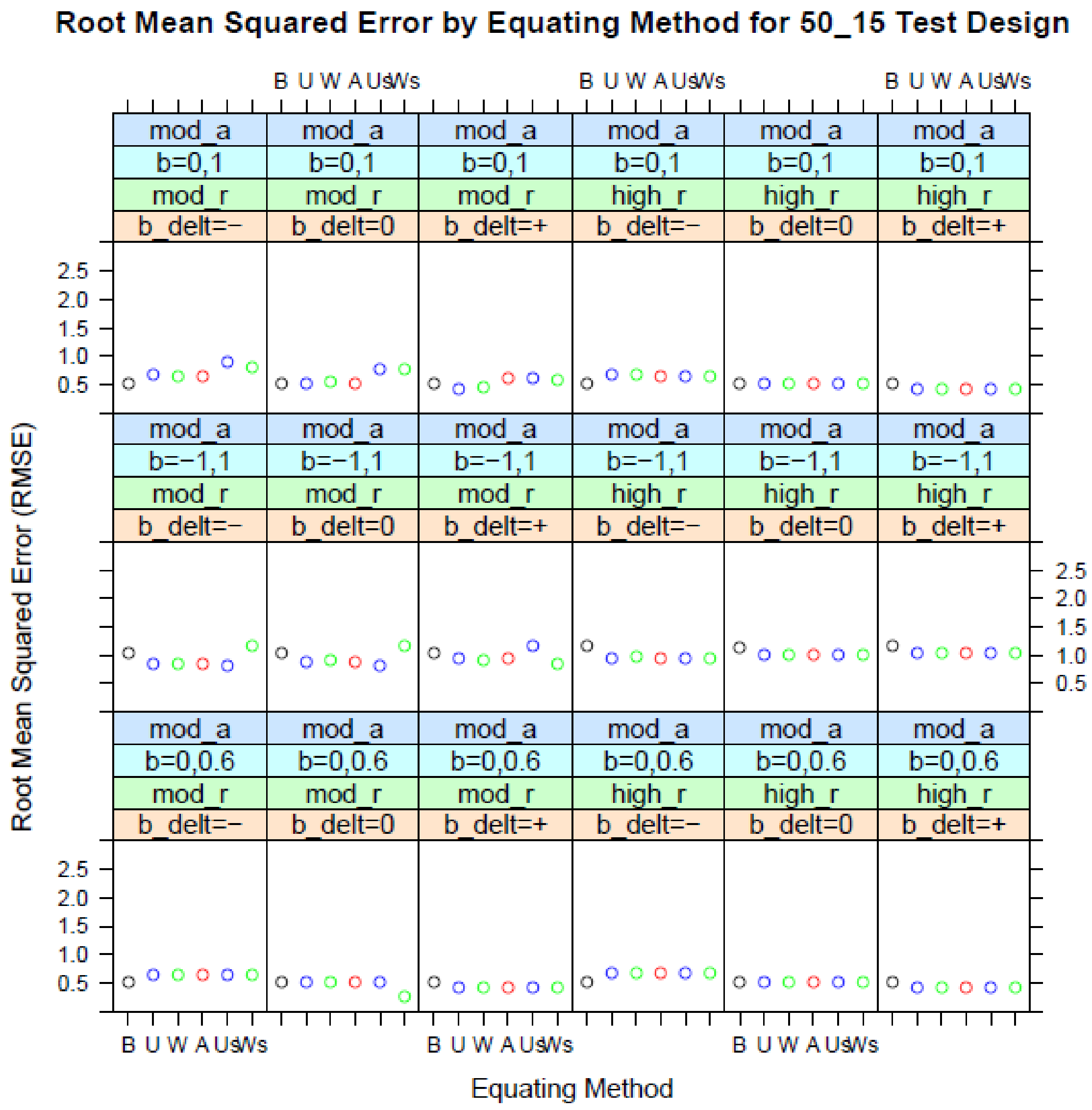


Figure 4.7a. Root mean squared error for moderate item discriminating conditions by equating method for 50_15 test design. (B = Base form, U = unweighted mean equating, W = weighted mean equating, A = Anchor Item Calibration, U$_s$ = stabilized unweighted mean equating with stabilization and W$_s$ = stabilized weighted mean equating with stabilization)

More importantly, the observations that relate change in mean anchor-item difficulty levels to RMSE that were made for 50_10 test design hold for this design. Precisely, when all conditions are held constant and mean anchor-item difficulty change (b_delt) varies, RMSE decreases as mean anchor difficult change becomes positive, i.e., from negative *b_delt* to positive *b_delt*. Therefore, RMSE is highest for negative *b_delt* and lowest for positive *b_delt* when all other study conditions remain constant. Within study conditions, the effect of using different equating methods is not noticeable. Thus, no precision gain in using stabilized equating over fixed equating methods.

In contrast, conditions with high discriminating items (high_a), exhibit similar patterns as those shown by conditions with low discriminating items discussed above. As Figure 4.7b indicates, RMSE is between 0.5 and 1 for all conditions with moderate mean difficulty and reasonable or constricted difficulty variability, b ~ (0, 1) or b ~ (0, 0.6). Similar to what was observed with low discriminating conditions, high discriminating conditions with low mean item difficulty for the whole test and reasonable variability, b ~ (-1, 1) exhibit higher RMSE values which range from 1 to about 1.25. The main difference between the two discriminating levels is that RMSE variability for high item discriminating and, b ~ (0, 1) conditions are higher than their moderate discriminating item-conditions counterparts for different equating methods. In addition, within study conditions, the effect of level of correlation on RMSE variability for the five equating methods is noticeable. Overall, study conditions composed of moderate correlation (mod_r), high item discrimination and, b ~ (0, 1) or, b ~ (0, 0.6) conditions show more differences (variability) in RMSE for the different equating methods. All other effects that apply to moderate item discriminating conditions also apply for high item discriminating conditions for this test design.

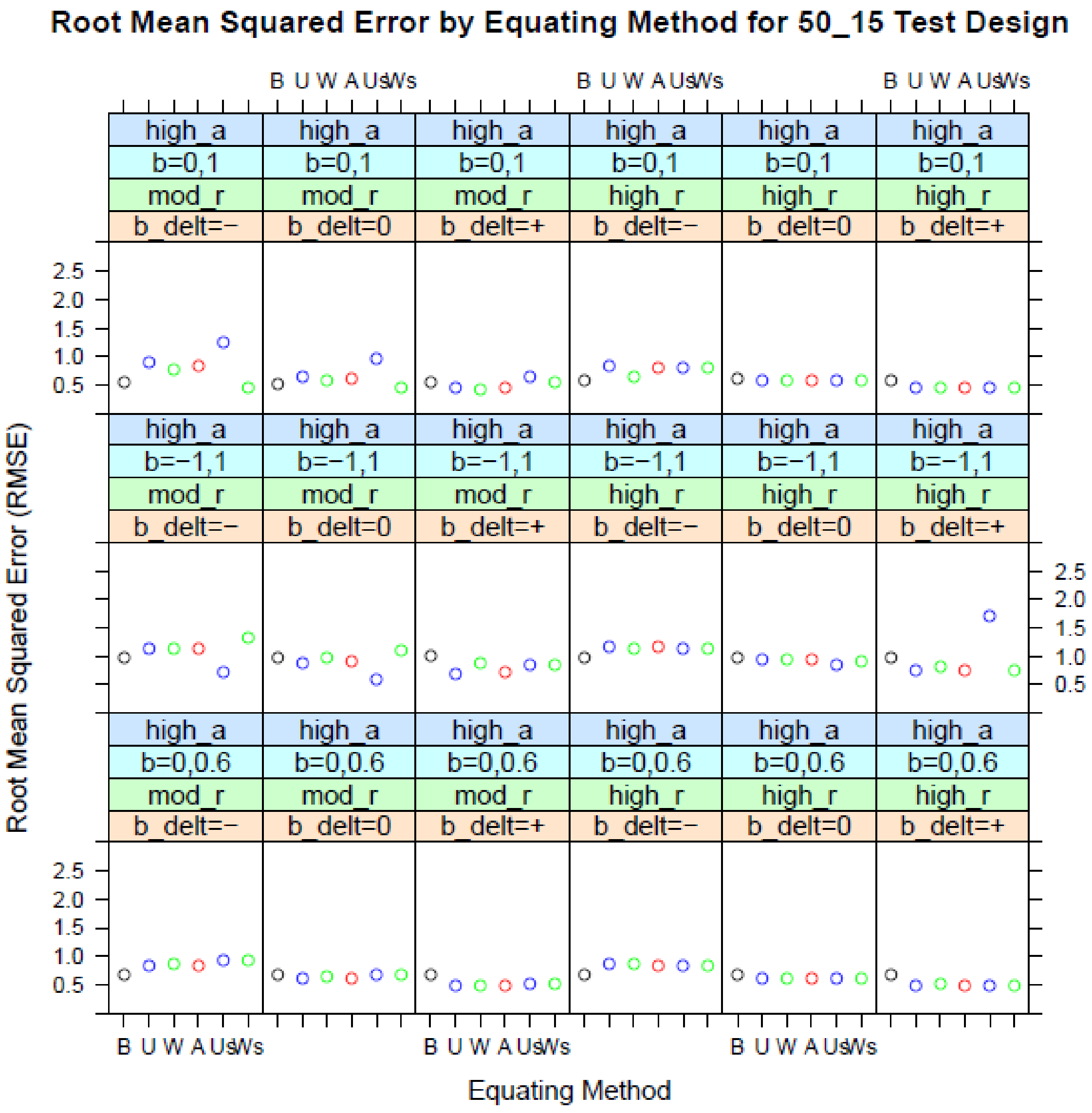


Figure 4.7b. Root mean squared error for high item discriminating conditions by equating method for 50_15 test design. (B = Base form, U = unweighted mean equating, W = weighted mean equating, A = Anchor Item Calibration, U_s = stabilized unweighted mean equating with stabilization and W_s = stabilized weighted mean equating with stabilization)

Bias and RMSE for 100_20 Test Design

The 100_20 test design consists of 100 items in all and 20 of these items are anchor items. Figures 29-34 in Appendix C show the different amounts of bias for different equating methods and study conditions. As noted from the graphs, the amount of bias is about zero for all test conditions where mean item difficulty for the whole test is moderate with reasonable or constricted variability, b ~ (0, 1) or b ~ (0, 0.6). In most cases (especially when b_delt < 0), mean bias is slightly above zero indicating that the Rasch ability estimates are overestimated. Also, there is a decreasing trend in bias when change in mean difficulty for anchor items on alternate forms changes from negative to positive (from b_delt < 0 to b_delt > 0). This implies that when all other study conditions remain constant and b_delt varies, the amount of bias is always at the lowest level for positive b_delt and highest for negative b_delt for all similar study conditions.

However, the largest amounts of bias were observed for all conditions where the mean item difficulty for the entire test is low with reasonable variability, b ~ (-1, 1) and item discrimination is high (high_a). Under these conditions, mean bias values are about 1.0 or slightly greater than 1.0. On the other hand, less discriminating conditions coupled with low mean item difficulty for the whole test and reasonable variability, b ~ (-1, 1) generally have negative bias values on average—on the order of magnitude of -.3 to -.4—but with more variability near the center of the population score distribution (referring to the length of the boxes in Figure 30 in Appendix C). The implication is that the true abilities are underestimated in this case. In addition, the decreasing bias trends that occur with changes in mean anchor-item difficulty on alternate forms corresponding increase for higher discriminating test conditions than for lower discriminating conditions.

Fortunately, from an operational perspective (and based on this study) the five equating methods generally seem to be performing in an equivalent manner in terms of bias. However, consistent with 50_10 and 50_15 test designs, levels of correlation between difficulty estimates of anchor-items on alternate forms and level of item discrimination determine the differences in mean bias for the five equating methods. In general, for b ~ (0, 1) and b ~ (0, 0.6), the five equating methods are susceptible to differences in mean bias when moderately correlated anchor item conditions coupled with high item discriminating conditions prevail. However, the differences in bias for the equating methods within such study conditions do not lead to any tangible results that can lead one to conclude that one equating method is more or less effective than the others, or that stabilization is better/worse than fixed equating. In short, all things being equal, the operationally least complex method may be as good as the most complicated method.

Figure 4.8a shows the amount of root mean squared error (RMSE) for various study conditions by equating method for the 100_20 test design when moderate item discriminating conditions are used. RMSE is between 0.4 and 0.9 for all conditions with moderate mean difficulty and reasonable or constricted difficulty variability, b ~ (0, 1) or b ~ (0, 0.6). However, for similar conditions that differ only when mean item difficulty for the whole test is low and the variability is reasonable, b ~ (-1, 1), RMSE is about 1. In addition, the effect of level of correlation is noticeable in that conditions with moderate correlation show more differences in RMSE for the different equating methods within study conditions.

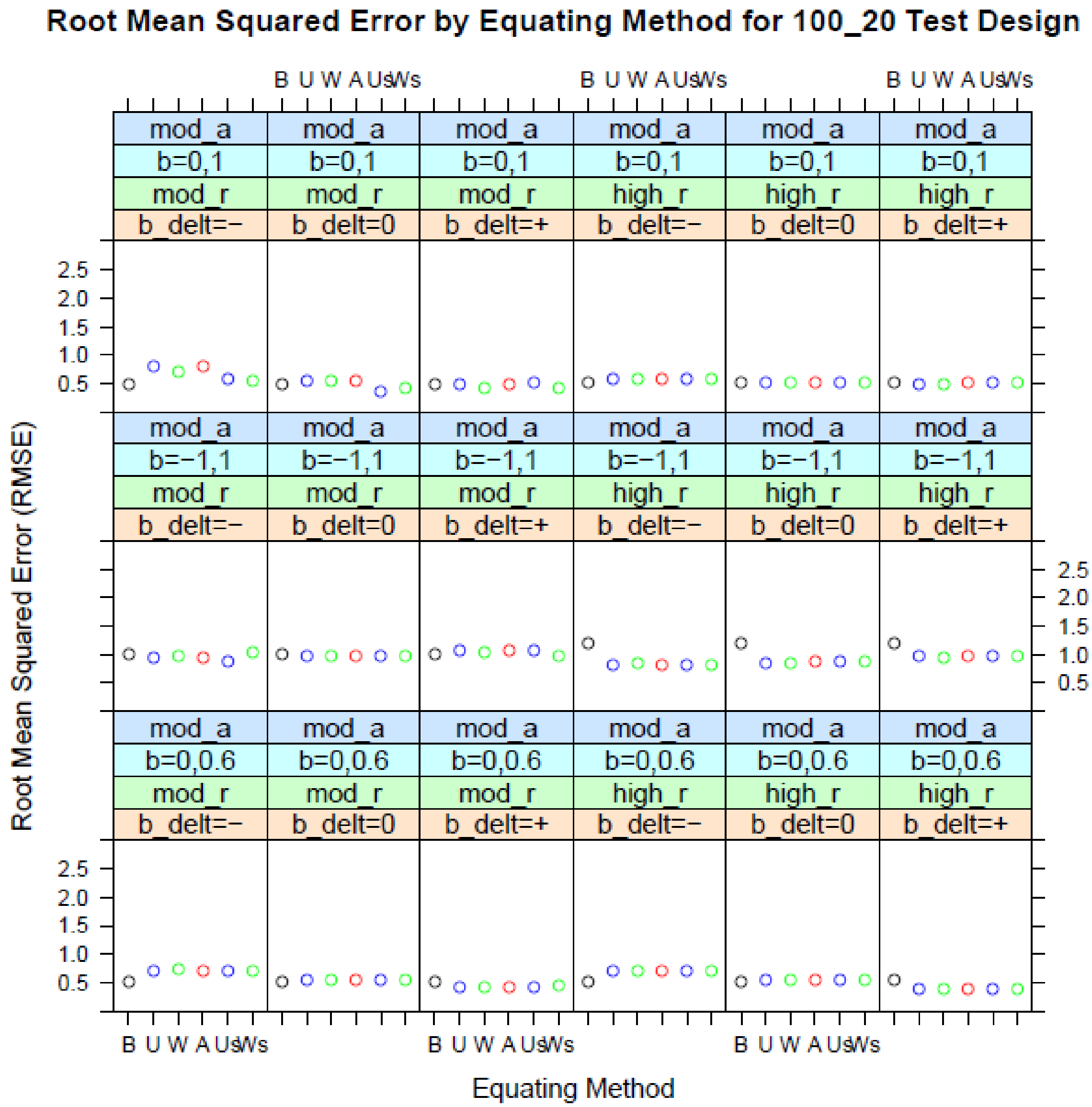


Figure 4.8a. Root mean squared error for moderate item discriminating conditions by equating method for 100_20 test design. (B = Base form, U = unweighted mean equating, W = weighted mean equating, A = Anchor Item Calibration, U_s = stabilized unweighted mean equating with stabilization and W_s = stabilized weighted mean equating with stabilization)

Furthermore, as was the case with the 50_10 and 50_15 test designs, when all conditions are held constant and mean anchor-item difficulty change (b_delt) is manipulated, RMSE decreases as mean anchor difficult change becomes positive, i.e., from negative b_delt to positive b_delt. Therefore, RMSE is highest for negative b_delt and lowest for positive b_delt when all other study conditions remain constant. Again, within study conditions, the effect of using different equating methods is not prevalent. Thus, no precision gain in using stabilized equating over fixed equating methods.

Moreover, when high item discrimination conditions are used, similar findings to those of low item discrimination conditions results. As Figure 4.8b indicates, RMSE is between 0.5 and 1.2 for all conditions with moderate mean difficulty and reasonable or constricted difficulty variability, b ~ (0, 1) or b ~ (0, 0.6). Higher values of about 1 to 1.2 are observed for high discriminating conditions with low mean item difficulty for the whole test and reasonable variability, b ~ (-1, 1). As was noted in the other test designs discussed before, RMSE differences among the equating methods manifest. Specifically, RMSE variability among equating methods for similar study conditions are higher for high item discriminating conditions than moderate item-discriminating conditions. In addition, within study conditions, the effect of level of correlation on RMSE variability for the five equating methods is noticeable.

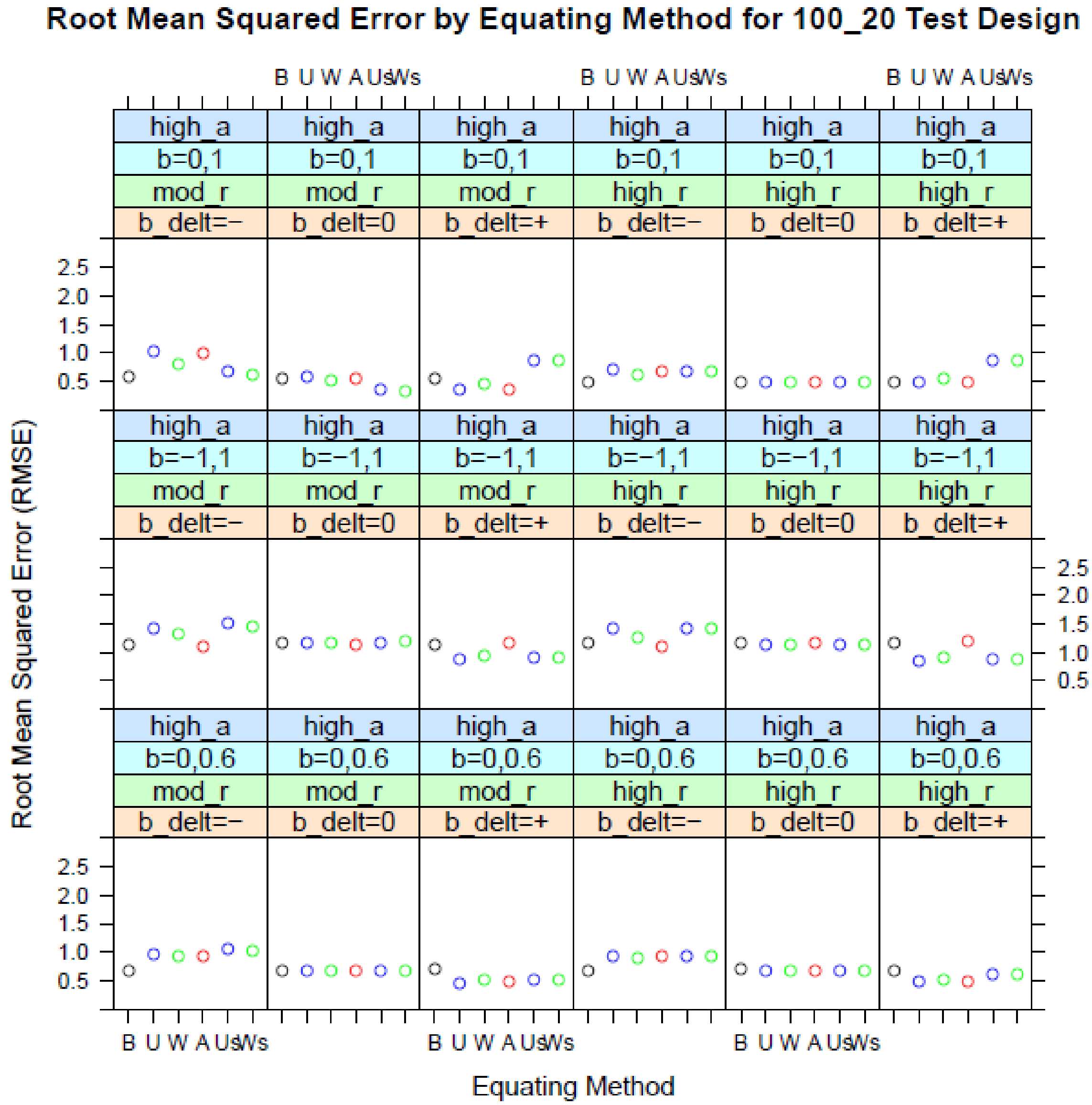


Figure 4.8b. Root mean squared error for high item discriminating conditions by equating method for 100_20 test design. (B = Base form, U = unweighted mean equating, W = weighted mean equating, A = Anchor Item Calibration, U_s = stabilized unweighted mean equating with stabilization and W_s = stabilized weighted mean equating with stabilization)

Bias and RMSE for 100_30 Test Design

The 100_30 test design consists of 100 items in all and 30 of these items are anchor items. Figures 35-40 in Appendix C show the different amounts of bias for different equating methods and study conditions. Clearly, the amount of bias is about zero for all test conditions where mean item difficulty for the whole test is moderate with reasonable or constricted variability, b ~ (0, 1) or b ~ (0, 0.6). In most cases though (especially where b_delt < 0), mean bias is slightly above zero indicating that the true ability estimates are overestimated. Also, just like the other three test designs already discussed, there is a decreasing trend in bias when change in mean difficulty for anchor items on alternate forms changes from negative to positive (from b_delt < 0 to b_delt > 0). This implies that when all other study conditions remain constant and b_delt varies, the amount of bias is always at lowest level (closer to zero) for positive b_delt and highest for negative b_delt (overestimated) for all similar study conditions.

Consistent with the other three test designs already discussed, the largest amounts of bias are observed for all conditions where the test is, on average, very easy, nonetheless with substantial variability in the item difficulties, b ~ (-1, 1) and where the average item discrimination is high. Under these conditions, the true abilities are overestimated and the average bias is about +1.0. As for less discriminating conditions coupled with low mean item difficulty for the whole test and reasonable variability, b ~ (-1, 1), negative bias values are generally observed on average (about - .2) but with more variability for the middle 50 percent bias values (indicated by longer box plots in Figure 36 in Appendix C). Ability estimates are underestimated in this case. In addition, for b ~

(-1, 1) conditions, there is also a decreasing bias trend that when changes in mean anchor-item difficulty on alternate forms increase.

Moreover, the three equating methods with two treatments for UME and WME seem to be performing in more or less the same way in terms of bias within study conditions. However, consistent with 50_10, 50_15, and 100_20 test designs, levels of correlation between difficulty estimates of anchor-items on alternate forms and level of item discrimination determine the differences in mean bias for the Rasch equating methods. It is clear that for conditions where mean item difficulty for the whole test are moderate with reasonable or constricted variability, that is, b ~ (0, 1) or b ~ (0, 0.6), the incidence of moderately correlated anchor items coupled with high item discriminating conditions may lead to larger differences in mean bias across all equating methods. However, the differences in bias for the equating methods within these study conditions do not lead to any tangible results that can lead one to conclude that one equating method is more effective than the others, or that stabilization is better/worse than fixed item (versus iteratively stabilized Rasch anchor-item) equating.

Figure 4.9a shows the amount of root mean squared error (RMSE) for various study conditions by equating method for the 100_30 test design when moderate item discriminating conditions are used. RMSE values are between 0.25 and 0.75 for all conditions with moderate mean difficulty and reasonable or constricted difficulty variability, b ~ (0, 1) or b ~ (0, 0.6). However, for similar conditions that differ only when mean item difficulty for the whole test is low and the variability is reasonable, b ~ (-1, 1), RMSE values are about 1.

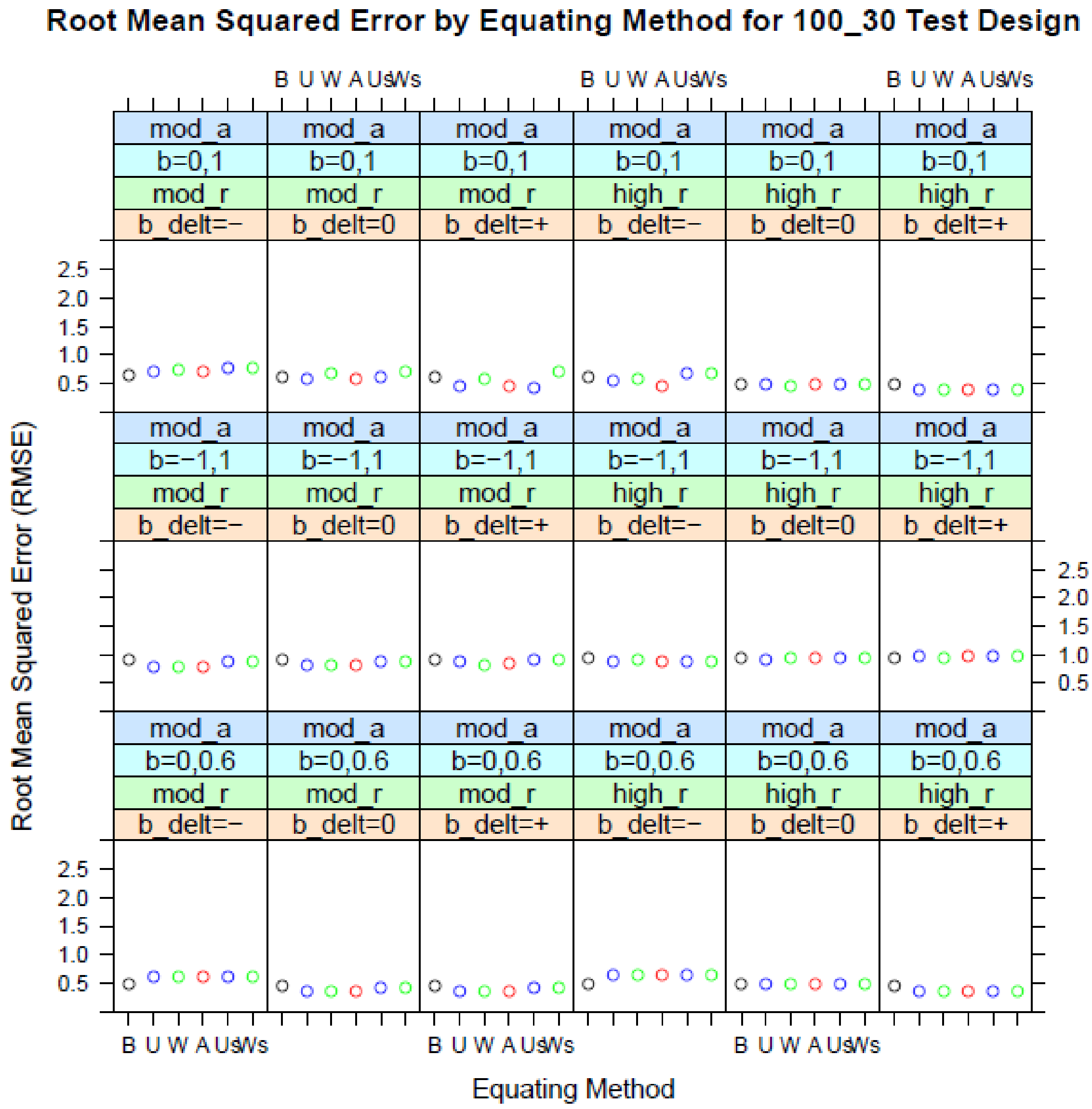


Figure 4.9a. Root mean squared error for moderate item discriminating conditions by equating method for 100_30 test design. (B = Base form, U = unweighted mean equating, W = weighted mean equating, A = Anchor Item Calibration, U_s = stabilized unweighted mean equating with stabilization and W_s = stabilized weighted mean equating with stabilization)

As was the case with the 50_10, 50_15 and 100_20 test designs, when all conditions are held constant and mean anchor-item difficulty change (b_delt) is manipulated, RMSE decreases as mean anchor difficult change becomes positive, i.e., from negative b_delt to positive b_delt. Therefore, RMSE is highest for negative b_delt and lowest for positive b_delt when all other study conditions remain constant.

In addition, the effect of level of correlation between the difficulty values of anchor-items is noticeable. In general, conditions with moderate correlation levels have higher differences in RMSE values than higher correlation conditions for different equating methods. Again, within study conditions, the effect of using different equating methods is not prevalent. Thus, no precision gain in using stabilized equating over fixed equating methods.

In contrast, when high item discrimination conditions prevail, as is depicted by Figure 4.9b , RMSE values are between about 0.4 and 1.25 for all conditions with moderate mean difficulty and reasonable or constricted difficulty variability, b ~ (0, 1) or b ~ (0, 0.6). Higher values of about 1 are consistently observed for high discriminating conditions with low mean item difficulty for the whole test and reasonable variability, b ~ (-1, 1). As was noted in the other three test designs discussed before, within study conditions, moderate correlation conditions and moderate mean difficulty and reasonable or constricted difficulty variability, b ~ (0, 1) or b ~ (0, 0.6) conditions exhibit higher differences in RMSE values for the Rasch equating methods. Also, RMSE variability among equating methods among similar study conditions are higher for high item discriminating conditions than moderate item-discriminating conditions. Therefore, as has been noted in similar conditions for the other test designs discussed earlier, moderate correlation conditions coupled with high item discriminating conditions lead to higher differences in RMSE for the different equating methods within study conditions.

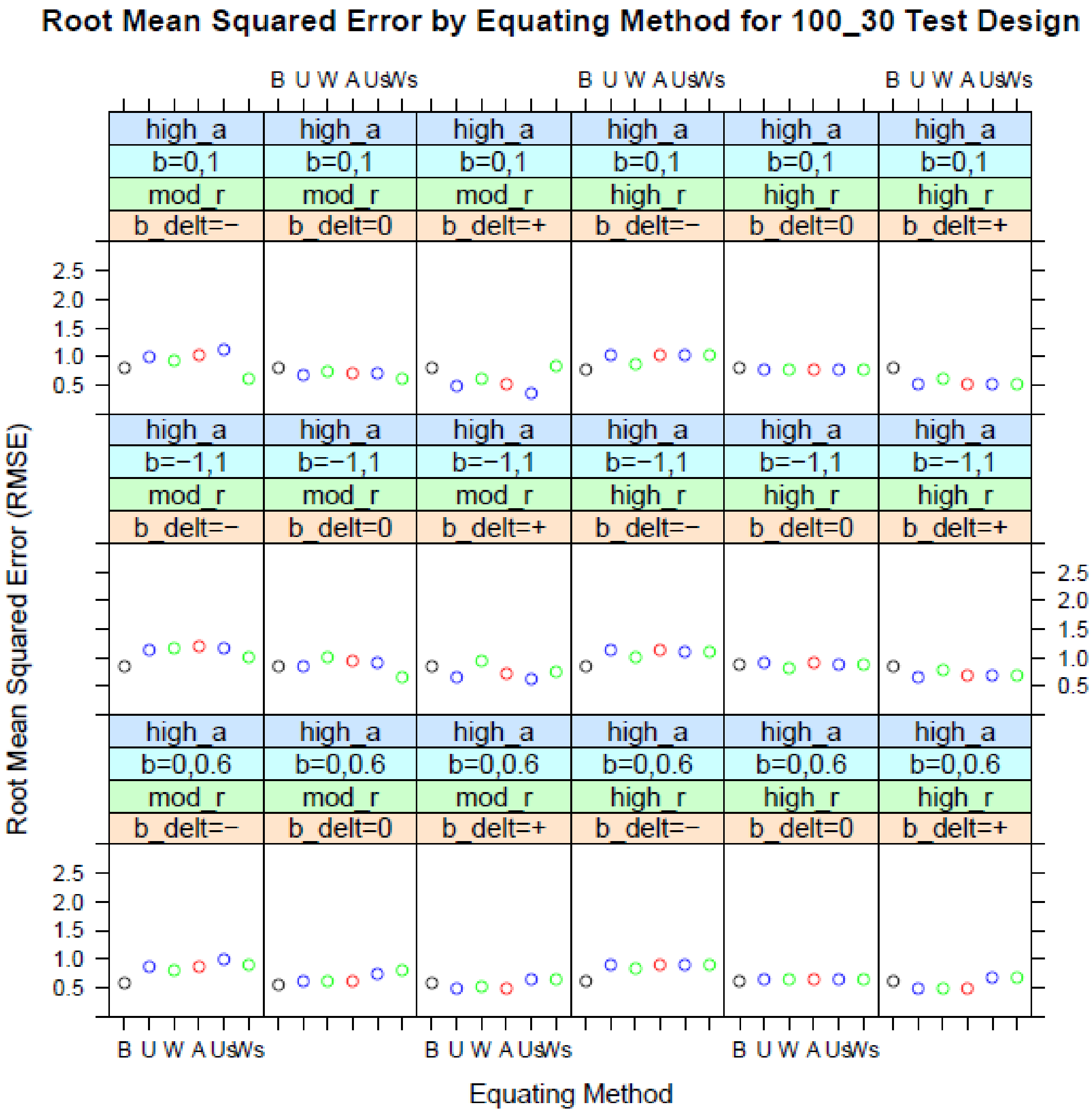


Figure 4.9b. Root mean squared error for high item discriminating conditions by equating method for 100_30 test design. (B = Base form, U = unweighted mean equating, W = weighted mean equating, A = Anchor Item Calibration, U_s = stabilized unweighted mean equating with stabilization and W_s = stabilized weighted mean equating with stabilization)

In conclusion, RMSE values for high item discriminating conditions neither provide enough evidence to suggest that anchor item stabilization lead to better or worse equating in comparison to fixed item equating nor lead us to conclude that at least one equating method is different from the other equating methods.

Summary of Bias and RMSE Results across Test Designs

For the four test designs, the ratio of anchor items to total number of items on the test is either .2 or .3. As expected, the similarities in structural designs (anchor-test ratio) and composition (study conditions considered) result into similar findings across test designs for the most part. Both bias and RMSE values across the four test designs show that all conditions where mean item difficulty for the whole test is moderate with reasonable or constricted variability i.e., b ~ (0, 1) or b ~ (0, 0.6) have low bias and RMSE values. On the other hand, high item discriminating conditions with low mean item difficulty for the whole test and reasonable variability, b ~ (-1, 1), have high bias and RMSE values. Ability values tend to be overestimated in such cases. However, moderate item discriminating conditions with low mean item difficulty for the whole test and reasonable variability, b ~ (-1, 1), have negative bias values.

As well as the four test designs showing similar patterns in the amounts of bias and RMSE for comparable study conditions, there is a decreasing trend in bias and RMSE for all variability and difficulty conditions when change in mean difficulty for anchor items on alternate forms changes from negative to positive (from b_delt < 0 to b_delt > 0). Therefore, when all other study conditions remain constant and b_delt varies, the amount of bias is always at the lowest level (closer to zero) for positive b_delt and highest for negative b_delt, for all similar study conditions.

Beyond similar amounts and similar trends for bias and RMSE for similar study conditions, the four test designs agree that there is not enough evidence to suggest that one equating method is better than the other four methods, or that stabilized equating is better than fixed equating for the three equating methods with two treatments for UME and WME. Although differences in bias and RMSE exist for the equating methods, especially with all conditions with high item discrimination (high_a) and moderate correlation (mod_a), the differences are random and do not lead to any tangible conclusions.

However, the differences in test lengths seem to have no profound impact on bias and RMSE values. Bias and RMSE values are roughly the same for similar study conditions for both shorter and longer test designs.

Classification Consistency

This section of results reports on the effects of using Rasch model estimates on the accuracy of classifying examinees into their original categories when the data was first generated. Largely, classification consistency and bias are similar since they both address the question of fit between generating and estimating models. In order to address the research question on whether the conditions of this study affected different ability examinees in different ways thoroughly, classification consistency results are summarized by test design. The effects of various study conditions are discussed for each of the three ability levels (below proficiency, just proficient and advanced) within each test design. For each test design, comparisons are also drawn among the three ability levels to determine the effect of study conditions and equating methods on classification rates. Finally, across test design comparisons for the three ability levels are made for similar ability levels.

Classification Consistencies within Test Designs

50_10 Test Design, Below Proficiency Ability Level

As discussed earlier, this test design comprises 50 items in all, 10 of which are anchor items. As is shown in Figures 41-44 in Appendix D, the effect on correct percent classification rate due to level of change in difficulty is very conspicuous. Generally, high item discriminating conditions lead to higher correct classification rates than moderate item discriminating conditions. Correct classification rates are mostly between 60 and 80 percent for most conditions where mean item difficulty for the whole test is moderate with reasonable or constricted variability, b ~ (0, 1) or b ~ (0, 0.6) and moderate item discriminating conditions. For similar conditions as discussed above that only differ in discriminating conditions, i.e., conditions with high item discriminations show improved correct classification rates, which are for the most part above 80 percent. However, correct classification rates drop when mean item difficulty for the whole test is low with reasonable variability, b ~ (-1, 1), which are mostly between lower 30s to about 50 percent for moderate discrimination conditions and above 60 percent for high discrimination conditions (see Figures 41 to 44 in Appendix D). As was the case with bias and RMSE, correct classification percentages improve with increasing difficulty change (b_delt) for anchor items on alternate forms for all conditions.

As for the different equating methods, it appears that within study conditions, all equating methods lead to similar percentages of correct classifications. Therefore, for this group of test takers, treating anchor items as stabilized versus fixed anchor item treatment does not result in higher correct classifications for all study conditions.

50_10 Test Design, Just Proficient Ability Level

As shown in Figures 45-48 in Appendix D, moderate item discriminating conditions lead to higher classification rates than high item discriminating conditions .Generally, correct classification rates are mostly between 80 and 90 percent for most moderate discrimination conditions. However, correct classification rates drop to about 50 - 70 percent for most high item discriminating conditions. In addition, there is no difference in correct percent classifications for the three difficulty-variability levels for similar study conditions. For this group of test takers, there is no increase in correct classification rate when change in mean difficulty for the anchor items on alternate forms (b_delt) is positive compared to negative difficulty change, a trend that was observed for below proficient examinees.

In terms of equating methods, there appears to be no advantage in using one method of equating over the other methods for all conditions. In addition, it seems that treating unweighted and weighted mean equating as stabilized or unstabilized does not make a difference.

50_10 Test Design, Advanced Ability Level

In general, as is shown in Figures 49-52 in Appendix D, percent correct classifications are very high. Most of the classification rates are between 80 percent and 100 percent. Percent classifications are about 100 percent when high discriminating conditions exist. In addition, all three difficulty-variability levels show no differences in correct percent classifications for similar study conditions. Further, classification rates fall to about 60 percent for moderate discrimination and b_delta > 0 conditions (see Figures 49 and 51).

In terms of equating methods, there seems to be no advantages in using one method of equating over the other four methods. All the equating methods appear to be working equally well as depicted from percent classification rates.

Comparisons across Ability Levels within 50_10 Test Design

The most notable difference among the three ability levels within this test design is that advanced ability-level examinees have the highest percent correct classification rates compared to below proficient and just proficient ability-level examinees for similar study conditions. Below proficient ability-level examinees show the lowest percent correct classification rates compared to the advanced and just proficient levels for similar study conditions (Tables 13-17 in Appendix B). While an increasing trend in percent classification rate due to mean anchor-item difficulty change on alternate forms is noticeable for all study conditions for below proficient levels, the trend seems to be nonexistent for just proficient levels and advanced levels. It seems there is a ceiling effect for such a trend for the already high percent classification rates for advanced ability-level examinees. Generally, for below proficient examinees, high item discriminating conditions lead to higher classification rates than moderate item discriminating conditions while the opposite is true for just proficient examinees.

In terms of choice of equating methods, it is clear that for all ability levels, no equating method is better than the other methods. Specifically, no equating method seems to have real impact on percent classification rates for below proficient, just proficient, and advanced ability-level examinees.

50_15 Test Design, Below Proficiency Ability Level

As discussed earlier, this test design comprises 50 items in all, 15 of which are anchor items. Generally, high item discriminating conditions lead to higher correct classification rates than moderate item discriminating conditions. As is shown in Figures 53-56 in Appendix D, classification rates are between upper 40s and 60 percent for all conditions with moderate item discrimination. Under high discriminating conditions, correct classification rates are between 60 and 80 percent. However, for both item discriminating conditions, correct classification rates are higher where mean item difficulty for the whole test is moderate with reasonable or constricted variability, b ~ (0, 1) or b ~ (0, 0.6) than correct classifications for conditions with low mean item difficulty for the whole test and reasonable variability, b ~ (-1, 1). In general, correct classification rates improve with increasing mean anchor-item difficulty change (b_delt) for anchor-items on alternate forms for all equating methods when all other conditions are held constant (i.e., as *b_delt* changes from negative to positive).

It is difficult to determine the most adequate equating method for this proficiency level. All equating methods are equally effective in terms of percent correct classification rates. Further, the evidence with respect to percent correct classification rates does not uphold the use of stabilized equating over fixed equating methods and vice versa.

50_15 Test Design, Just Proficient Ability Level

In general, high item discriminating conditions have lower correct classification rates than moderate item discriminating conditions. Correct classification rates for just proficient ability-level examinees are between lower 70s and 90 percent for all moderate item discrimination conditions and drop to between 60 and 70 percent for similar study conditions that differ only in item discrimination, i.e., when high item discrimination conditions exist (see Figures 57–60). The lowest correct classification rates for this proficiency levels are observed when low mean item difficulty with reasonable variability, b ~ (-1, 1), moderate correlation (mod_r) and high item discrimination conditions exist (see Figures 58 where classification rates are between 40 and 50 percent). In addition, there is an increase in correct classification rates with increasing mean difficulty change for anchor items on alternate forms (b_delt) for all study conditions.

Overall, there is no advantage in using one equating method over the other methods for this ability-level within study conditions. There is no difference in correct classification rates for all Rasch equating methods. Therefore, the use of stabilized or fixed treatments does not have any added advantage.

50_15 Test Design, Advanced Ability Level

As is shown in Figures 61-64 in Appendix D, for all study conditions, percent correct classifications are very high. Most of the classification rates are between 90% and 100%. Percent classifications are about 100 percent when high discriminating conditions exist. However, study conditions with moderate correlation between anchor item difficulties (mod_r), moderate mean item difficulty for the whole test and constricted difficulty variability, b ~ (0, 1) or b ~ (0, 0.6), and positive difficulty change for anchor items (b_delt) conditions have about 80 percent correct classification rate. This amount of correct classification rate for these study conditions seem to deviate from the usual high consistent rates that are a trademark of advanced level examinees but are not a major source of concern. In addition, when low mean item difficulty with reasonable variability, b ~ (-1, 1) and moderate discrimination conditions exist, percent correct classifications are lower than the other two difficulty-variability levels for similar conditions.

As for equating methods, there seems to be no advantages in using one method of equating over the other since all the methods appear to be working equally well as depicted from percent classifications. By extension, these results indicate that there are no differences to support the use of stabilized equating methods over unstabilized or vice-versa.

Comparisons across Ability Levels within 50_15 Test Design

In general, all equating methods indicate that advanced ability level examinees have the highest correct classification rates. Below proficient examinees, on the other hand have the lowest classification rates for similar study conditions (see Tables 18-22 in Appendix B). In addition, for advanced level examinees, study conditions with moderate discrimination (mod_a), moderate mean item ability for the whole test with constricted variability (b ~ 0, 0.6), and positive change in mean anchor-item difficulty (b_delt > 0) showed lower correct classification rates of about 80 percent (see Figure 61). Similarly, when low mean item difficulty with reasonable variability, b ~ (-1, 1) and moderate discrimination exist, correct classification rates slightly above 80 percent (see Figures 61 and 63). Although these are not low classification rates, they are not in line with the other high correct classification rates for advanced ability-level study conditions.

Study conditions where mean item difficulty for the whole test is low and with reasonable variability, b ~ (-1, 1) indicate that such difficulty level leads to very low classification rates for below proficient examinees (when moderate item discrimination conditions exist) and just proficient examinees (when high item discrimination conditions exist). Also, for below proficient examinees and just proficient examinees there is an increase in classification rates with increasing mean difficulty change for anchor items on alternate forms (*b_delt*).

Overall, in terms of choice of equating methods, it is clear that for all ability levels, no single equating method works better than the other methods with respect to correct classification rates.

100_20 Test Design, Below Proficiency Ability Level

As has been stated already, this test design has 100 items in all and 20 of these items are anchor items. In general, moderate item discrimination conditions have lower correct classification rates (between 30 to 40 percent) than high item discrimination conditions where classification rates are about 50 to 80 percent for all study conditions as shown in Figures 65-68 in Appendix D. In addition, as is the case with the other test designs discussed earlier, there is an increasing trend in classification rates with increasing mean difficulty change for anchor items on alternate forms (*b_delt*). The lowest correct classification rates are observed when moderate item discrimination and b~ (-1, 1) conditions exist. In these cases, correct classification rates fall to about 20 percent (see Figure 65).

It is difficult to determine the most adequate equating method for this test design and proficiency level. All equating methods are equally effective in terms of percent correct classification rates. Further, the evidence with respect to classification rates does not uphold the use of stabilized equating over fixed equating methods and vice versa.

100_20 Test Design, Just Proficient Ability Level

As has already been observed with the other two test designs, high item discriminating conditions have lower classification rates than moderate item discriminating conditions for this ability level. Percent correct classification rates for just proficient ability-level examinees are between lower 70s and 90 percent (for moderate item discrimination conditions), shown in Figures 69-71 in Appendix D. However, correct classification rates are lower for high item discrimination conditions (see Figures 70 and 72 in Appendix D), about 50 to 60 percent with the lowest conditions showing an average correct classification rate of about 40 percent and is observed when high item discrimination conditions with b~ (-1, 1) exist. Further, for the most part, classification rates increase with increasing difficulty change for anchor items on alternate forms (*b_delt*).

Overall, no equating method works better than the other four methods within study conditions for this ability-level. The impact of method of equating on classification rates is non-evident. It follows therefore that use of stabilized equating has no advantage over fixed item equating or vice versa.

100_20 Test Design, Advanced Ability Level

As is shown in Figures 73-76 in Appendix D, for all study conditions, percent correct classifications are very high. Most of the classification rates are between 95 and 100 percent. In addition, for all difficulty-variability levels, when mean difficulty change for anchor items is positive (b_delt > 0), correct classification rates drop to about 80 to 90 percent. These study conditions seem to depart from the usual high consistent rates that are a trademark of advanced ability-level examinees. The expectation is that these harder items challenge some of the advanced examinees thereby causing variability in examinee responses, which in turn lead to lower classification rates. The decrease in classification rates are however minimal and are not a major source of concern.

The ceiling effect in percent correct classifications makes it difficult to determine an equating method that shows more equating adequacy than the other methods. The three methods of equating with two treatments on UME and WME appear to be working equally well. In addition, the effect of stabilizing anchor items is not noticeable.

Comparisons across Ability Levels within 100_20 Test Design

As expected, close inspections among similar study conditions indicate that advanced ability-level examinees have the highest correct classification rates whereas below proficient ability-level examinees have the lowest classification rates (see Tables 23-27 in Appendix B). Study conditions where mean item difficulty for the whole test is low and with reasonable variability, b ~ (-1, 1) indicate that such difficulty level leads to very low classification rates for below proficient examinees (when moderate item discrimination conditions exist) and just proficient examinees (when high item discrimination conditions exist). Also, for below proficient examinees and just proficient examinees there is an increase in correct classification rates with increasing mean difficulty change for anchor items on alternate forms (*b_delt*). Again, there are clear patterns that suggest that some advanced examinees encounter problems with difficult items for study conditions with moderate discriminating items and positive mean anchor-item difficult changes.

Overall, for all the three ability groups, there seems to be no equating method that works better than the other methods. Again, the effect of stabilization, if it exists, is not observable.

100_30 Test Design, Below Proficiency Ability Level

At this ability level for the 100_30 test design, which has 100 items (including 30 anchor-items), percent correct classification rates are generally low. As is shown in Figures 77-80 in Appendix D, in general, moderate item discrimination conditions have lower correct classification rates (between 20 to 60 percent) than high item discrimination conditions where classification rates are about 50 to 70 percent for all study conditions. In addition, similar to other test designs discussed earlier, there is an increasing trend in classification rates with increasing mean difficulty change for anchor items on alternate forms (*b_delt*). The lowest correct classification rates are observed when moderate item discrimination and b~ (-1, 1) conditions exist. In these cases, correct classification rates fall to about 20 to 30 percent.

As for the equating methods, there are no classification rate trends caused by differences in equating methods that translate into a definite interpretable pattern, one that will enable the possibility of drawing the conclusion that at least one classification method leads to high classification rates for this ability level. For the same reason, it is inconclusive whether anchor item treatment (stabilization or fixed), leads to higher classification rates than the other.

100_30 Test Design, Just Proficient Ability Level

For this ability level, when moderate item discrimination conditions are used, percent correct classification rates are generally between 80s and upper 90s for all levels of item difficulty and variability conditions, i.e., b ~ (0, 1), b ~ (0, 0.6) and b ~ (-1, 1). Alternatively, when high item discrimination conditions are used, correct classification rates are lower, generally between 50s and 60s for all levels of item difficulty, and variability conditions (see Figures 81-84 in Appendix D). Moreover, there is an increasing percent classification trend as change in mean anchor item difficulty increases for all study conditions. The lowest correct classification rates are observed when high item discrimination conditions with b~ (-1, 1) exist. In such cases correct classification rates drop to about 40 percent.

It appears no equating method supersedes the other methods in terms of having higher percent classification rates. As a result, the choice of equating method, stabilized or fixed seems to have no impact on classification rates.

100_30 Test Design, Advanced Ability Level

As is shown in Figures 85–88 in Appendix D, correct classification rates are very high for all study conditions. Most of the classification rates are between 90 and 100 percent. In addition, for all difficult-variability levels, study conditions with moderate discrimination and conditions with positive difficulty change for anchor items (*b_delt*> 0) have about 80% to 90% classification rates (see Figures 85 and 87). For these study conditions, the increasing difficulties for anchor items for moderate difficult items lead to incorrect responses by some examinees. This in turn creates variability in performance for the advanced ability-level examinees. As stated in other similar circumstances, the decrease in correct classification rates are however minimal and are not a major source of concern.

With such high percent correct classification rates, it is difficult to determine an equating method that shows more adequacy than the other methods. The three methods of equating appear to be working equally effective in terms of classification consistency rates. Therefore, the effect of stabilized or fixed methods for treating anchor items is not obvious.

Comparisons across Ability Levels within 100_30 Test Design

Similar study conditions across different ability-levels within the 100_30 test design indicate that advanced ability-level examinees have the highest correct classification rates while below proficient examinees have the lowest correct classification rates (Tables 28–32 in Appendix B). Study conditions where mean item difficulty for the whole test is low, and with reasonable variability, b ~ (-1, 1) have the lowest classification rates for below proficient examinees (when moderate discrimination conditions exist) and just proficient examinees (when high discrimination conditions exist). Also, for below proficient examinees and just proficient examinees there is an increase in classification rates with increasing mean difficulty change for anchor items on alternate forms (b_delt).

In addition, advanced ability-level examinees encounter issues with lower percent classification rates. For all difficult-variability levels, study conditions with moderate discrimination and conditions with positive difficulty change for anchor items (*b_delt*> 0) have lower than expected high classification rates.

It appears no equating method supersedes the other methods in terms of having higher percent classification rates for all three ability levels. As a result, the choice of equating method, stabilized or fixed seems to have no impact on classification rates.

Summary of Classification Results across Test Designs

Despite differences in the total number of test and anchor items for the four test designs, similar results exist across test designs for similar ability levels and comparable study conditions. In the first place, advanced ability-level examinees for all four test designs have high classification rates. This is expected because most of the items are easy for advanced ability-level examinees. However, when difficulty increases, not all examinees correctly respond to the given items. This leads to lower classification rates although not very low to cause concern.

Another similarity among the four test designs is that study conditions where mean item difficulty for the whole test is low and with reasonable variability, b ~ (-1, 1) lead to very low classification rates for below proficient examinees (when moderate item discrimination conditions exist) and for just proficient examinees (when high item discrimination conditions exist). The low correct classification rates for below proficient examinees become even more conspicuous with longer tests. For example, the 100_30 test design shows lower percent correct classification rates than the 50_10 test design for similar study conditions. This implies that with longer tests, researchers are able to isolate study conditions that lead to worst misfit more precisely than they can with shorter tests. This in turn allows test developers to concentrate on those conditions that improve fit for lower ability examinees.

Also, for below proficient examinees and just proficient examinees there is an increase in correct classification rates with increasing mean difficulty change for anchor items on alternate forms (*b_delt*). As items become harder, low ability examinees have a low chance of giving a correct response. Therefore, most examinees classified as failing during generation are more likely to be classified as failing when Rasch equating methods are used. This pattern is not observable for advanced level examinees as there is no room for increase for correct classification rates, which are already high.

From an equating methods point of view, all methods of equating seemed to be equally adequate. No definite pattern could be discerned from the performance of equating methods with respect to percent correct classification rates for all the four test designs. Stated precisely, using percent classifications, one cannot tell the advantage of using unweighted mean equating over weighted mean equating or vice versa. Even the use of anchor item calibration for equating purposes does not exhibit any clear advantages to unweighted or weighted mean equating. In addition, the advantage of using stabilized analyses for weighted mean equating and unweighted mean equating over fixed item equating does not clearly manifest when correct classification rates are used.

CHAPTER V

CONCLUSIONS AND DISCUSSION

Overview of the Chapter

The purpose of this study was to examine the impacts of item difficulty changes for anchor test items on Rasch equating results and the quality of score estimates for performance related decisions. Real and simulated data were used. While the real data investigated the effects of item position on examinees' responses, the simulated data investigated the effects of item parameter changes as caused by context effects in general through manipulation of whole test variables and anchor test variables. Analysis of real and simulated data showed that lack of item parameter invariance affects the quality of equating resulting into poor decisions based on performance of examinees. This chapter specifically summarizes the implications of findings in reference to the research questions that were raised in chapter one. In conclusion, a discussion on the limitations and future studies follows.

Summary of Findings

The first research question sought to investigate the magnitude and direction of item difficulty changes and conditional probability changes in relation to serial item position changes at different proficiency levels on the ability scale. Overall, there was sufficient evidence to conclude that the position of an item determines the difficulty of an item and the probability of an examinee's correct response if it were given in a different position. However, in line with Wise et al. (1989) findings, this study also finds that the effects of item position changes are more pronounced for low ability examinees than high ability examinees. The study also shows high variability in probability differences for low ability examinees than that of high ability examinees. This implies that high ability examinees are more likely to be more consistent in their responses despite item position changing.

The second research question and its associated sub questions aimed at examining the adequacy of equating for the different conditions manipulated in the simulation study using bias and RMSE. Although procedures were carried out to ensure that the high bias and RMSE values were not a recentering artifact, it was apparently clear that the worst study conditions for the four test designs were those where the whole test comprised easy items with reasonable difficulty variability, b~ (-1,1). Items at this level of difficulty were tailored specifically for low ability examinees. As expected, easy items will discriminate among low ability examinees while high ability examinees will consistently perform better on such items. In general, easy items cannot help make the distinctions among high ability examinees because they are more likely to get all the items correct. Such variability in responses for low ability examinees will introduce some statistical noise, which will imply that the estimating models will not fully reflect the generating model hence the observed discrepancies as shown by large amount of bias. In addition, under these worst conditions, all equating methods performed equally poor. In general, no equating method performed better than the other methods as they all show high bias and RMSE values.

The best study conditions for the four test designs were those that had moderate difficult items with reasonable or constricted difficulty variability. These conditions registered low bias and RMSE values. With the increase in item difficulty, it is expected that low ability examinees will incorrectly respond to items that are higher than their ability level. On the other hand, high ability examinees will get all the items correct because they are still lower than their ability level. Low bias and RMSE values for this study conditions imply that the expectations are met resulting in almost similar ability values for both the generating and estimating models. Further, all the equating methods performed equally well so much so that it was not possible to isolate an equating method which was better than the other methods.

The above findings seem to be consistent among the four test designs. In addition, there is a notable difference in RMSE values due to differences in test lengths. RMSE values decreased as test length increased. The effect of having low RMSE value was eminent in the decreasing trend in bias as change in anchor-item difficulty increases. The differences in amount of bias that exist among equating methods within study conditions, although small, were lower for longer tests than for shorter tests.

The third research question compared the two item treatments (stabilization versus fixed item formats) to determine whether such treatment of anchor items improved the quality of equating. Results show that there was not enough evidence to conclude that stabilization was better than fixed or vice versa. For the four test designs, especially when mean test difficulty was moderate, most items had similar standard errors (about 0.03), which imply that all items were equally likely to be pruned. As a result, many items remained as anchor items and the effect of stabilization could not be sufficiently established for these good study conditions. Alternatively, the easy test conditions provided poor study conditions to differentiate the effects of stabilized and fixed equating methods. Although most items were pruned for most study conditions where easy items were used, these study conditions were not good that any equating differences that would have occurred might not be meaningful. The poor study conditions and effects pruning are more likely to bring about a confounded interpretation on the effects of stabilization. It will be worthy to investigate the effects of stabilization when study conditions are good, e.g., for all conditions with moderate mean test-difficulty.

From the ongoing discussion, it follows that this research could not establish whether pruning longer/shorter anchor items had a major role on the effects of stabilization. This is based on the premise that when longer anchor tests are used, there are still enough anchor items left in the common item anchor set after stabilization to maintain at least 20% anchor item presence purported in psychometric literature (Kolen & Brennan, 2004). Alternatively, the use of shorter anchor tests leads to very few items remaining in the common item anchor set, and in many cases, less than the recommended proportion required which ultimately leads to inadequate equating.

This research however found that when conducting weighted mean equating with a stability analysis, the stability analysis prunes out most of the items from the common item anchor set supposedly because the weights destabilize the iterative process, causing the weighted means and provisional equating constant to be erratic. This problem seems to occur when the standard errors of the item difficulty estimates differ between the base and alternate forms. Equal standard errors (or no weighting) work fine in a stability analysis. Therefore, if the goal is to have a stable equating function, the existence of unequal standard errors may cause problems for a stability analysis that uses weighted mean equating.

The need for bias free testing situations was reflected in the fourth research question and its associated sub questions. Consistent with the findings in the first research question, all test designs found that classification rates for low ability examinees were worse than classification rates for high ability examinees for similar study conditions. In general, study conditions that comprised easy items show low classification rates for below proficient and basic ability-level examinees whereas high ability examinees are not affected. As was discussed earlier, easy items differentiate low ability and basic ability groups of examinees. It is therefore not surprising to see low levels in classification rates for below proficient and basic ability-level examinees and no effect on advanced ability-level examinees for similar study conditions. On the other hand, all conditions with moderate mean difficulty show high classification rates. The classification rates for advanced level examinees are much higher than the classification rates for below proficient and basic examinees. It was also observed that some study conditions that comprised moderate mean difficulty for the whole test and positive changes in anchor item difficulty exhibited lower than expected classification rates. A possible explanation for this is that the increasing difficulty of the anchor items shifts the difficulty level of items from moderate status towards higher difficulty levels. This has the effect of making the test harder and essentially tailored to discriminate among high ability examinees. It will be interesting to investigate further, how the use of much harder items could affect high ability examinees.

In terms of equating methods, there is no evidence to suggest that at least one equating method led to higher correct classification rates than the other methods of equating for all test designs under similar study conditions. Even when best study conditions are used, the equating methods worked equally fine in terms of percent correct classifications. The same was true for the worst study conditions where no equating method could be isolated as the better equating method for classification purposes. These findings are in line with the findings using bias and RMSE. However, the findings on correct classification rates offer more information in connecting the effects of item invariance on different ability levels on the score scale, a premise that could not be fully investigated using bias and RMSE.

Finally, correct classification rate findings were consistent over various test lengths. All test designs exhibited similar patterns of correct classification rates for similar study conditions

Practical Implications of Results

As Leary and Dorans (1985) and other researchers have expressed, context effect-issues are common in psychometric literature. Therefore, to some degree, all tests deal with context effects. The existence of context effects imply that item parameter invariance principle becomes under threat. Meyers et al. (2009) have highlighted that the effect of context effects are reflected in item difficulty and the resultant equating. However, very few studies have demonstrated how changes in item parameters affect equating. This research is an attempt for a renewed interest in this field of research. The findings from this research serve as a reminder to test developers to institute and or fortify research on two fronts. First, test developers need to continually check that the quality of items they use are of high standards. As was the case with this study where poor conditions prevailed, rewriting of items to maintain high standards becomes necessary.

In the real world, true equating functions are difficult to establish. In fact, as Kolen and Brennan (2004) put it, the ideal equating likely has never been conducted in practice. However, it is in the interest of every test developer to administer parallel forms that make the examinees to be indifferent on the form they want to take. If the different forms have almost the same difficulty, equating becomes easier. To this end, the second area that test developers need to fine tune is the precision of measurement among different groups of examinees. In conclusion, test developers must acknowledge that there is imprecision in every measurement. However, they have to do something about it. The following quote from Standards for educational and psychological testing could not be expressed any better: “Although it is never possible to achieve perfect accuracy in describing an individual’s performance, efforts need to be made to minimize errors in estimating individual scores or in classifying individuals in pass/fail or admit/reject categories” (p. 139).

Limitations and Future Studies

As has been established already, examinees’ item responding behavior is affected by many test context effects. These effects include the location of an item within a test

(Davey & Lee, 2010; Hill, 2008; Meyers et al., 2009), wording, content, format (Kingston & Dorans, 1984; Zwick, 1991) and specific features of other items that surround it (Davis & Ferdous, 2005; Haladyna, 1992) and many other factors. The current research has not addressed all test context effects but has attempted to investigate changes in item positions as an example of test context effects. Further, because the effects of such changes manifest on item difficulty, item difficulty changes were discussed at length in this research. In addition, only three Rasch equating/linking methods with two treatments on anchor items were used for this investigation. Since there are a lot of equating methods that many test developers use other than the ones reported herein, future studies may expand this study to include other equating methods and more factors such as other item formats (other than multiple choice), to investigate whether the effects will be consistent with the findings in this study.

Item position is just one facet of test context effects and this research investigated the effects of item position in relation to item parameter invariance-principle or specifically changes in item parameter estimates. In this study, item shift distances were defined by the number of positions an item moved from its original position to another position in different or same section of the test. It did not matter which section an item moved from and to— initial administration and where it migrated. However, assuming that same shift distances have the same effect irrespective of the sections where items are shifting may obscure the whole picture. The reality is that it is more likely that the same amounts of shift distances from one section of the test to another (e.g., from the middle of the test to the beginning of the test or from the end of the test to the middle of the test) may not have the same effects. Future lines of research will incorporate the effect of section to acknowledge the likely effects that shifting from one section of a test to another may have on item difficulty.

Other lines of research that emerge from this study include isolating the type of ratios for item-difficulty standard errors between the base and the alternate forms when using weighted equating stability analysis. This as reported earlier, operates on the principle that unequal standard errors for the base and alternate forms does not work fine for stability analysis using weighted mean equating as the weighted means and provisional equating constants become erratic. Future simulation studies will therefore address different ratios for item-difficulty standard errors between the base and the alternate forms to determine the permissible range of standard errors that could work to avoid pruning all the items.

Finally, the use of simulations allowed for direct manipulation of the nature and extent of item difficulty changes to investigate (a) conditions that lead to worst/best testing situations, (b) the effect of item parameter changes on equating, and (c) the impact of item difficulty changes on score accuracy and decision/classification accuracy. For this research, a number of specific study conditions, which included test variables, anchor item test variables and examinee characteristics were set and controlled in the simulation study. In addition, as was pointed out in chapter two, the use of 3 PL model entails that c_i, a lower asymptote parameter that is associated with noisy response patterns that is typical of low ability examinees due to guessing is embraced for data generation. The effect of the use of a guessing parameter as De Ayala (2009) describes, is that it makes some examinees at the lower end of the ability scale to have a higher probability of providing a correct response. This in turn makes their ability appear more than what they actually should be. Further, an attempt to create some statistical noise, which is typical of real data, was made using the 3PL model for data generation and 1 PL for estimations. It is possible that the observed impacts on lower ability examinees in this study are likely an artifact of 3 PL use for data generation. Although these attempts to mimic real data situations seem sufficient, they should not be perceived as the norm on what to expect when real data is used. It is expected that when real data is used the situation will be more complex because there will be more factors involved that the present simulation study cannot fully address. It will be interesting to investigate the adequacy of equating for the equating methods used in this research with testing programs that use similar test designs as the ones used in this study.

REFERENCES

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational & Psychological Testing (US). (1999). *Standards for educational and psychological testing*. American Educational Research Association.

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* (2nd ed.). Washington, DC: American Council on Education.

Baker, F. B. (1985). *The basics of item response theory* (1st ed.). Portsmouth, NH: Heinemann.

Berger, V. F., Munz, D. C., Smouse, A. D., & Angelino, H. (1969). The effects of item difficulty sequencing and anxiety reaction type on aptitude test performance. *Journal of psychology*, *71*(2), 253-258.

Bergstrom, B., Gershon, R., & Lunz, M. (1994). *Computerized adaptive testing: exploring examinee response time using hierarchical linear modeling.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Brenner, M. H. (1964). Test difficulty, reliability, and discrimination as functions of item difficulty order. *Journal of Applied Psychology*, *48*(2), 98.

Birnbaum, A. (1968). Some latent train models and their use in inferring an examinee's ability. *Statistical theories of mental test scores*, 395-479.

Bridgeman, B., & Cline, F. (2002). *Fairness issues in computer adaptive tests with strict time limits.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

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

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

Cohen, J., Jiang, T., & Yu, P. (2009). *Information-weighted linking constants*. American Institutes for Research.

Cook, L., & Eignor D. (1991). An NCME instructional module on IRT equating methods. *Educational Measurement: Issues and Practice, 10*(3), 37–45.

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

Davey, T., & Lee, Y. H. (2010). *An examination of context effects in linear test forms with effects pretested in a random context*. Paper presented at the annual meeting of the National Council on Measurement in Education, Denver, Colorado.

Davis, J., & Ferdous, A. (2005). *Using item difficulty and item position to measure test fatigue*. Paper presented at the Annual meeting of the American Educational Research Association, Montreal, Quebec.

De Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, NY: The Guilford Press.

Debeer, D., & Janssen, R. (2012). *Modeling item-position effects within an IRT-framework.* Paper presented at the annual meeting of the National Council on Measurement in Education, Vancouver, British Columbia, Canada.

De Boeck, P., & Wilson, M. (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York, NY: Springer.

DeMars, C. (2010). *Item response theory*: *Understanding statistics measurement*. New York, NY: Oxford.

Dorans, N. J., & Lawrence, I. M. (1990). Checking the statistical equivalence of nearly identical test editions. *Applied Measurement in Education, 3,* 245–254.

Eignor, D. R., & Stocking, M. (1986). *An Investigation of possible causes for the inadequacy of IRT pre-equating* (ETS RR-86-14). Princeton, NJ: Educational Testing service.

Ferguson, G. A. (1942). Item selection by the constant process. *Psychometrika*, *7*(1), 19-29.

Flaugher, R. L., Melton, R. S., & Myers, C. T. (1968). Item rearrangement under typical test conditions. *Educational and Psychological Measurement*, *28*(3), 813-824.

Folk, V. G., & Smith, R. L. (2002). Models for delivery of CBTs. In C. N. Mills, M. T. Potenza, J. J. Fremer, & W. C. Ward (Eds.), *Computer based testing: Building the foundation for future assessments* (pp. 41–66). Mahwah, NJ: Lawrence Erlbaum Associates.

Graybill, F. A., & Deal, R. D. (1959). Combining unbiased estimators. *Biometrics, 15,* 543–550.

Haladyna, T. (1992). Context-dependent item sets. *Educational Measurement: Issues and Practice, 11*(1), 21–25.

Haertel, E. (2004). *The behavior of linking items in test equating.* CSE Report 630. CRESST.

Hambleton, R. K., & Rodgers, J. (1995). Item bias review. *Practical Assessment, Research, and Evaluation, 4*(6). Retrieved February 10, 2012, from http://PAREonline.net/getvn.asp?v=4&n=6

Hambleton, R. K., & Swaminathan, H. (1984). *Item response theory: Principles and applications* (Vol. 7). Kluwer Academic Pub.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer.

Hambleton, R. K., & Traub, R. E. (1974). The effects of item order on test performance and stress. *Journal of Experimental Education, 43,* 40–46.

Harris, D. (1991). Effects of passage and item scrambling on equating relationships. *Applied Psychological Measurement, 15*(3), 247–256.

He, W., Gao, R., & Ruan, C. (2009). *Does pre-equating work? An investigation into pre-equated testlet-based college placement exam using post administration data.* Paper presented at the annual meeting of the National Council on Measurement in Education (NCME), San Diego, California.

Hill, R. (2008). *Using P-value statistic to determine the believability of equating results*. Paper presented at the National Conference on student assessment, Orlando, Florida.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. *Test validity*, 129-145.

Huck, S. W., & Bowers, N. D. (1972). Item difficulty level and sequence effects in multiple-choice achievement tests. *Journal of Educational Measurement*, *9*(2), 105-111.

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, S. P. (1981, October). *The effects of time limits, item sequence, and question format on applicant performance on the California Bar Examination: A report submitted to the Committee of Bar Examiners of the State Bar of California and the National Conference of Bar Examiners.* Report #81-7. Retrieved from http://www.seaphe.org/pdf/past-bar-research/The_Effects_of_Time_Limits,_Item_Sequence.pdf

Klosner, N. C., & Gellman, E. K. (1973). The effect of item arrangement of classroom test performance: Implications for content validity. *Educational and Psychological Measurement,* 33, 413- 418.

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

Kubinger, K. D. (2008). On the revival of the Rasch model-based LLTM: From constructing tests using item generating rules for measuring item administration effects. *Psychology Science Quarterly, 50*(3), 311–327.

Kubinger, K. D. (2009). Applications of the linear logistic test model in psychometric research. *Educational and Psychological Measurement, 69*(2), 232–244.

Lawley, D. N. (1943). The application of the maximum likelihood method to factor analysis. *British Journal of Psychology. General Section*, *33*(3), 172-175.

Leary, L., & Dorans, N. (1982). *The effects of item rearrangement on test performance: A review of the literature.* (ETS Research Report RR-82-30). Princeton, NJ: Educational Testing Service.

Leary, L., & Dorans, N. (1985). Implications for altering the context in which test items appear: A historic perspective on an immediate concern. *Review of Educational Research, 55,* 387–413.

Linacre, J. M., & Wright, B. D. (2011). WINSTEPS (Version 3.73. 0).

Lord, F. M. (1952). The relation of the reliability of multiple-choice tests to the distribution of item difficulties. *Psychometrika*, *17*(2), 181-194.

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

Luecht, R.M. (2012a). 1PLIWgtEqt Version 1.0 software manual, August 2012.

Luecht, R.M. (2012b). *GEN3PL equating version of response generation software*.

Marso, R. N. (1970). Test item arrangement, testing time, and performance. *Journal of Educational Measurement*, *7*(2), 113-118.

Meyers, J. L., Kong, X. J., & McClarty, K. L. (2008). *An investigation of changes in item parameter estimates for items re-field tested.* Paper presented at the Annual Meeting of the American Educational Research Association, New York.

Meyers, J. L., Miller, G. E., & Way, W. D. (2009). Item position and item difficulty change in an IRT-based common item equating design. *Applied Measurement in Education, 22*(1), 38–60.

Meyers, J. L., Murphy, S., Goodman, J., & Turhan, A. (2012). *The impact of item position changes on item parameters and common item equating results under the 3PL model*. Paper presented at the annual meeting of the National Council on Measurement in Education, Vancouver, British Columbia, Canada.

Mills, C. N., Potenza, M. T., Fremer, J. J., & Ward, W. C. (Eds.). (2002). *Computer-based testing: Building the foundation for future assessments*. Routledge.

Mislevy, R. (1992). *Linking educational assessments: Concepts issues, methods, and prospects.* Princeton, NJ: ETS Policy Information Center.

Mollenkopf, W. G. (1950). *Prediction of second-year grade point averages at the US naval post graduate school*. Princeton, NJ: Educational Testing Service.

Monk, J. J., & Stallings, W. M. (1970). Effects of item order on test scores. *The Journal of Educational Research*, 463-465.

Muraki, E., Hombo, C. M., & Lee, Y. W. (2000). Equating and linking of performance assessments. *Applied Psychological Measurement*, *24*(4), 325-337.

Murphy, S., Little, I., Kirkpatrick, R., Fan, M., & Lin C. H. (2010). *The impact of different anchor stability methods on equating results and student performance*. Paper presented at the annual meeting of the National Council on Measurement in Education, Denver, Colorado.

Petersen, N. S., Cook, L. L., & Stocking, M. L. (1983). IRT versus conventional equating methods: A comparative study of scale stability. *Journal of Educational Statistics, 8,* 137–156.

Petersen, N. S. (2008). A discussion of population invariance of equating. *Applied Psychological Measurement*, *32*(1), 98-101.

Perrone, M. (2006). Differential item functioning and item bias: Critical considerations in test fairness. *Columbia University Working Papers in TESOL and Applied Linguistics*, *6*(2). Retrieved February 10, 2012, from http://www.tc.columbia.edu/tesolawebjournal

Plake, B. (1984). Social and technical issues in testing: Implications for test construction and usage. *Buros-Nebraska Symposium on measurement and Testing, 1*. Hillsdale, NJ: Lawrence Erlbaum.

Plake, B. S., Ansorge, C. J., Parker, C. S., & Lowry, S. R. (1982). Effects of item arrangement, knowledge of arrangement test anxiety and sex on test performance. *Journal of Educational Measurement*, *19*(1), 49-57.

Plake, B. S., Thompson, P. A., & Lowry, S. (1981). Effect of item arrangement, knowledge of arrangement, and test anxiety on two scoring methods. *The Journal of Experimental Educational*, 214-219.

Pommerich, M., & Harris, D. J. (2003). *Context effects in pretesting: Impact on item statistics and examinee scores*. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago.

Rijmen, F., & De Boeck, P. (2002). The random weights linear logistic test model. *Applied Psychological Measurement, 26*, 271–285.

Roever, C. (2005). *"That's not fair!" Fairness, bias and differential item functioning in language testing*. Retrieved February 10, 2012, from the University of Hawai'i System Website: http://www2.hawaii.edu/~roever/brownbag.pdf

Rubin, L., & Mott, D. (1984). *The effect of the position of an item within a test on the item difficulty value.* Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

Sax, G., & Carr, A. (1962). An investigation of response sets on altered parallel forms. *Educational and Psychological Measurement t*, 22, 371-376.

Sax, G., & Cromack, T. R. (1966). The Effects of Various Forms of Item Arrangements on Test Performance. *Journal of Educational Measurement*, *3*(4), 309-311.

Sinharay, S., & Holland, P. (2006). *Choice of anchor test in equating.* (ETS Research Report RR-06-35). Princeton, NJ: Educational Testing Service.

Sinharay, S., & Holland, P. (2006). *The correlation between the scores of a test and an anchor test* (ETS Research Report RR-06-35). Princeton, NJ: Educational Testing Service.

Smouse, A. D., & Munz, D. C. (1968). Interaction effects of item-difficulty sequence and achievement-anxiety reaction on academic performance. *Journal of Educational Psychology, 59*(5), 370.

Smouse, A. D., & Munz, D. C. (1969). Item difficulty sequencing and response style: A follow-up analysis. *Educational and Psychological Measurement,* 29, 469-472.

Subkoviak, M. J., Mack, J. S., Ironson, G. H., & Craig, R. D. (1984). Empirical comparison of selected item bias detection procedures with bias manipulation. *Journal of Educational Measurement, 21*(1), 49–58.

Talento-Miller, E., & Guo F. (2009). Guess what? Score differences with rapid replies versus omissions on a computerized adaptive test. In D. J. Weiss (Ed.), *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing.* Retrieved 05/21/2012 from www.psych.umn.edu/psych.umn.edu/psylabs/ CATCentral/

Talento-Miller, E., Rudner, L., Han, C., & Guo F. (2012). *Analysis of pretest items by position*. Paper presented at the annual meeting of the National Council on Measurement in Education, Vancouver, British Columbia, Canada.

Towle, N. J., & Merrill, P. F. (1975). Effects of anxiety type and item-difficulty sequencing on mathematics test performance. *Journal of Educational Measurement, 12*(4), 241-249.

Van der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of modern item response theory*. New York, NY: Springer.

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

Wingersky, M. S., Cook, L. L., & Eignor, D. R. (1987). *Specifying the characteristics of linking items used for item response theory item calibration*. (ETS Research Report RR-87-24). Princeton, NJ: Educational Testing Service.

Wise, L., Chia, W., & Park, R. (1989). *Item position effects for test of work knowledge and arithmetic reasoning*. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco.

Wilson, M. (2005). *Constructing measures: An item response modeling approach.* Mahwah, NJ: Lawrence Erlbaum.

Whitely, E., & Dawis, R. (1976). The influence of test context on item difficulty. *Educational and Psychological Measurement, 36,* 329–337.

Wu, M. (2010). Measurement, sampling, and equating errors in large-scale assessments. *Educational Measurement: Issues and Practice, 29*(4), 15–27.

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.

Zumbo, B. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores.* Ottawa ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Zwick, R. (1991). Effects of item order and context of estimation on NAEP reading proficiency. *Educational Measurement: Issues and practice, 10,* 10–16.

Zwick, R. (1992). Statistical and Psychometric issues in the measurement of Educational Achievement Trends: Examples from the National Assessment of Educational Progress. *Journal of Educational and Behavioral Statistics, 17,* 205–218.

APPENDIX A

PROPORTIONS OF ANCHOR ITEMS TO TOTAL NUMBER OF ITEMS

Table 12

Proportions of Anchor Items to Total Number of Items after Stabilization for the Four Test Designs under Different Study Conditions

				Equating Method and Test Design							
Conditions				50_10		50_15		100_20		100_30	
a_ level	b/var_level	Correlation level	Anchor b- Change	UME	WME	UME	WME	UME	WME	UME	WME
Moderate	Normal	Moderate	-.25	.10	.08	.16	.14	.11	.11	.17	.15
			.00	.10	.08	.18	.18	.10	.10	.18	.17
			+.25	.12	.12	.18	.16	.06	.07	.15	.22
		High	-.25	.20	.20	.30	.30	.20	.20	.09	.09
			.00	.20	.20	.30	.30	.20	.20	.30	.30
			+.25	.20	.20	.30	.30	.20	.20	.30	.30
	Easy/ Reasonable	Moderate	-.25	.12	.12	.10	.12	.11	.08	.13	.13
			.00	.12	.12	.08	.12	.10	.10	.12	.12
			+.25	.10	.10	.10	.12	.11	.10	.12	.12
		High	-.25	.20	.20	.30	.30	.20	.20	.30	.29
			.00	.20	.20	.30	.30	.20	.20	.30	.30
			+.25	.20	.20	.30	.30	.20	.20	.30	.30
	Moderate/ Constricted	Moderate	-.25	.18	.18	.22	.22	.15	.15	.21	.21
			.00	.18	.18	.24	.24	.14	.14	.22	.22
			+.25	.18	.18	.22	.22	.14	.15	.23	.23
		High	-.25	.20	.20	.30	.30	.20	.20	.30	.30
			.00	.20	.20	.30	.30	.20	.20	.30	.30
			+.25	.20	.20	.30	.30	.20	.20	.30	.30

Table 12 (cont.)

				Equating Method and Test Design							
Conditions				50_10		50_15		100_20		100_30	
a_ level	b/var_level	Correlation level	Anchor b- Change	UME	WME	UME	WME	UME	WME	UME	WME
High	Normal	Moderate	-.25	.08	.08	.12	.08	.08	.08	.10	.05
			.00	.06	.08	.14	.08	.09	.09	.10	.10
			+.25	.10	.10	.14	.10	.07	.07	.09	.04
		High	-.25	.18	.18	.30	.30	.20	.20	.30	.28
			.00	.20	.20	.30	.30	.20	.20	.30	.30
			+.25	.20	.18	.30	.30	.09	.09	.30	.28
	Easy/ Reasonable	Moderate	-.25	.08	.06	.10	.10	.08	.08	.08	.05
			.00	.08	.08	.12	.10	.08	.08	.09	.07
			+.25	.06	.08	.10	.10	.09	.09	.08	.03
		High	-.25	.20	.18	.06	.06	.20	.19	.29	.27
			.00	.20	.20	.08	.06	.20	.20	.29	.29
			+.25	.20	.18	.08	.06	.20	.20	.30	.28
	Moderate/ Constricted	Moderate	-.25	.14	.10	.12	.12	.12	.11	.16	.16
			.00	.10	.10	.12	.12	.10	.11	.16	.14
			+.25	.14	.08	.12	.12	.11	.11	.15	.15
		High	-.25	.20	.20	.30	.30	.20	.20	.30	.30
			.00	.20	.20	.30	.30	.20	.20	.30	.30
			+.25	.20	.20	.30	.30	.20	.20	.30	.30

APPENDIX B

COMPARISONS ACROSS ABILITY LEVELS

Table 13

Number of Test Takers Correctly Classified for Unweighted Mean Equating Method for 50_10 Test Design

Test Characteristics					Difficulty_Variability Level and Equating Method					
					Moderate/ Reasonable		Easy/ Reasonable		Moderate/ Constrict	
Design	a_ Lvl	Cor_ Lvl	Anc_b change	Prof Level	3 PL Classif	UME$_F$ Classif	3 PL Classif	UME$_F$ Classif	3 PL Classif	UME$_F$ Classif
50_10	Mod	Mod	-.25	Bel_prof	1577	687	424	157	1547	618
				Basic	2771	2223	1281	1138	2777	1982
				Advanced	652	609	3295	2847	676	666
			.00	Bel_prof	1572	877	463	214	1595	897
				Basic	2763	2409	1332	1245	2711	2248
				Advanced	665	587	3205	2538	694	652
			+.25	Bel_prof	1540	1180	521	326	1570	1178
				Basic	2787	2498	1351	1296	2760	2375
				Advanced	673	474	3128	2003	670	568
		High	-.25	Bel_prof	1551	717	417	121	1585	616
				Basic	2774	2267	1054	987	2748	2087
				Advanced	675	627	3529	2819	667	640
			.00	Bel_prof	1560	1010	446	194	1551	790
				Basic	2774	2338	1088	1047	2786	2364
				Advanced	666	579	3466	2483	663	610

Table 13 (cont.)

Test Characteristics					Difficulty_Variability Level and Equating Method					
					Moderate/ Reasonable		Easy/ Reasonable		Moderate/ Constrict	
Design	a_ Lvl	Cor_ Lvl	Anc_b change	Prof Level	3 PL Classif	UME$_F$ Classif	3 PL Classif	UME$_F$ Classif	3 PL Classif	UME$_F$ Classif
50_10			+.25	Bel_prof	1542	1198	472	303	1553	1077
				Basic	2786	2507	1130	1070	2782	2435
				Advanced	672	460	3398	1887	665	562
	High	Mod	-.25	Bel_prof	1535	916	336	281	1581	959
				Basic	2792	1970	2028	1170	2759	1438
				Advanced	673	668	2636	2625	660	660
			.00	Bel_prof	1548	1358	345	311	1576	1277
				Basic	2787	2031	2143	1577	2765	1757
				Advanced	665	647	2512	2449	659	659
			+.25	Bel_prof	1517	1464	385	380	1543	1466
				Basic	2805	2067	2176	1655	2813	1942
				Advanced	678	605	2439	2194	644	637
		High	-.25	Bel_prof	1569	1070	281	205	1552	995
				Basic	2734	1832	1968	1017	2752	1509
				Advanced	697	697	2751	2747	696	696
			.00	Bel_prof	1514	1371	327	294	1532	1246
				Basic	2820	2044	2021	1458	2784	1765
				Advanced	666	663	2652	2592	684	682
			+.25	Bel_prof	1583	1537	352	347	1592	1517
				Basic	2736	1978	2056	1584	2714	1876
				Advanced	681	622	2592	2331	694	686

Table 14

Number of Test Takers Correctly Classified for Weighted Mean Equating Method for 50_10 Test Design

Test Characteristics					Difficulty_Variability Level and Equating Method					
					Moderate/ Reasonable		Easy/ Reasonable		Moderate/ Constrict	
			Anc_b		3 PL	WME$_F$	3 PL	WME$_F$	3 PL	WME$_F$
Design	a_ Lvl	Cor_ Lvl	change	Prof Level	Classif	Classif	Classif	Classif	Classif	Classif
50_10	Mod	Mod	-.25	Bel_prof	1577	953	424	157	1547	618
				Basic	2771	2449	1281	1138	2777	1982
				Advanced	652	527	3295	2847	676	666
			.00	Bel_prof	1572	1133	463	214	1595	897
				Basic	2763	2408	1332	1245	2711	2248
				Advanced	665	558	3205	2538	694	652
			+.25	Bel_prof	1540	1180	521	326	1570	1178
				Basic	2787	2498	1351	1277	2760	2375
				Advanced	673	474	3128	2238	670	568
		High	-.25	Bel_prof	1551	717	417	121	1585	616
				Basic	2774	2267	1054	987	2748	2087
				Advanced	675	627	3529	2819	667	640
			.00	Bel_prof	1560	1010	446	194	1551	790
				Basic	2774	2338	1088	1047	2786	2364
				Advanced	666	579	3466	2483	663	610
			+.25	Bel_prof	1542	1198	472	264	1553	1077
				Basic	2786	2507	1130	1080	2782	2435
				Advanced	672	460	3398	2150	665	562

Table 14 (cont.)

Test Characteristics					Difficulty_Variability Level and Equating Method					
					Moderate/ Reasonable		Easy/ Reasonable		Moderate/ Constrict	
Design	a_ Lvl	Cor_ Lvl	Anc_b change	Prof Level	3 PL Classif	WME$_F$ Classif	3 PL Classif	WME$_F$ Classif	3 PL Classif	WME$_F$ Classif
50_10	High	Mod	-.25	Bel_prof	1535	1165	336	255	1581	959
				Basic	2792	2057	2028	1034	2759	1438
				Advanced	673	661	2636	2631	660	660
			.00	Bel_prof	1548	1358	345	294	1576	1363
				Basic	2787	2031	2143	1434	2765	1711
				Advanced	665	647	2512	2478	659	659
			+.25	Bel_prof	1517	1464	385	368	1543	1416
				Basic	2805	2067	2176	1575	2813	1875
				Advanced	678	605	2439	2365	644	642
		High	-.25	Bel_prof	1569	1070	281	239	1552	1109
				Basic	2734	1832	1968	1320	2752	1487
				Advanced	697	697	2751	2710	696	696
			.00	Bel_prof	1514	1286	327	294	1532	1246
				Basic	2820	2131	2021	1304	2784	1765
				Advanced	666	663	2652	2627	684	682
			+.25	Bel_prof	1583	1464	352	335	1592	1487
				Basic	2736	1979	2056	1446	2714	1818
				Advanced	681	657	2592	2508	694	690

Table 15

Number of Test Takers Correctly Classified for Anchor Item Calibration Method for 50_10 Test Design

Test Characteristics					Difficulty_Variability Level and Equating Method					
					Moderate/ Reasonable		Easy/ Reasonable		Moderate/ Constrict	
Design	a_ Lvl	Cor_ Lvl	Anc_b change	Prof Level	3 PL Classif	AIC Classif	3 PL Classif	AIC Classif	3 PL Classif	AIC Classif
50_10	Mod	Mod	-.25	Bel_prof	1577	687	424	157	1547	618
				Basic	2771	2223	1281	1138	2777	1982
				Advanced	652	609	3295	2847	676	666
			.00	Bel_prof	1572	877	463	245	1595	897
				Basic	2763	2409	1332	1239	2711	2107
				Advanced	665	587	3205	2538	694	664
			+.25	Bel_prof	1540	1180	521	326	1570	1178
				Basic	2787	2498	1351	1277	2760	2375
				Advanced	673	474	3128	2238	670	568
		High	-.25	Bel_prof	1551	717	417	121	1585	507
				Basic	2774	2267	1054	987	2748	1906
				Advanced	675	627	3529	2819	667	657
			.00	Bel_prof	1560	880	446	194	1551	790
				Basic	2774	2385	1088	1047	2786	2364
				Advanced	666	579	3466	2483	663	610
			+.25	Bel_prof	1542	1198	472	303	1553	1077
				Basic	2786	2440	1130	1070	2782	2435
				Advanced	672	518	3398	1887	665	562

Table 15 (cont.)

Test Characteristics					Difficulty_Variability Level and Equating Method					
					Moderate/ Reasonable		Easy/ Reasonable		Moderate/ Constrict	
Design	a_ Lvl	Cor_ Lvl	Anc_b change	Prof Level	3 PL Classif	AIC Classif	3 PL Classif	AIC Classif	3 PL Classif	AIC Classif
50_10	High	Mod	-.25	Bel_prof	1535	916	336	255	1581	959
				Basic	2792	1970	2028	1034	2759	1438
				Advanced	673	668	2636	2631	660	660
			.00	Bel_prof	1548	1358	345	311	1576	1277
				Basic	2787	2031	2143	1577	2765	1757
				Advanced	665	647	2512	2449	659	659
			+.25	Bel_prof	1517	1464	385	380	1543	1466
				Basic	2805	2067	2176	1655	2813	1942
				Advanced	678	605	2439	2194	644	637
		High	-.25	Bel_prof	1569	1070	281	205	1552	995
				Basic	2734	1832	1968	1017	2752	1386
				Advanced	697	697	2751	2747	696	696
			.00	Bel_prof	1514	1371	327	294	1532	1246
				Basic	2820	2044	2021	1304	2784	1765
				Advanced	666	663	2652	2627	684	682
			+.25	Bel_prof	1583	1537	352	347	1592	1517
				Basic	2736	1978	2056	1584	2714	1876
				Advanced	681	622	2592	2331	694	686

Table 16

Number of Test Takers Correctly Classified for Unweighted Mean Equating Method with Stabilization for 50_10 Test Design

Test Characteristics					Difficulty_Variability Level and Equating Method					
					Moderate/ Reasonable		Easy/ Reasonable		Moderate/ Constrict	
Design	a_ Lvl	Cor_ Lvl	Anc_b change	Prof Level	3 PL Classif	UME_S Classif	3 PL Classif	UME_S Classif	3 PL Classif	UME_S Classif
50_10	Mod	Mod	-.25	Bel_prof	1577	553	424	157	1547	720
				Basic	2771	2067	1281	1138	2777	2118
				Advanced	652	627	3295	2847	676	656
			.00	Bel_prof	1572	877	463	214	1595	1036
				Basic	2763	2271	1332	1245	2711	2212
				Advanced	665	613	3205	2538	694	652
			+.25	Bel_prof	1540	1363	521	280	1570	1178
				Basic	2787	2297	1351	1294	2760	2375
				Advanced	673	403	3128	2238	670	568
		High	-.25	Bel_prof	1551	717	417	121	1585	507
				Basic	2774	2267	1054	987	2748	1906
				Advanced	675	627	3529	2819	667	657
			.00	Bel_prof	1560	880	446	194	1551	790
				Basic	2774	2385	1088	1047	2786	2364
				Advanced	666	579	3466	2483	663	610
			+.25	Bel_prof	1542	1198	472	303	1553	1077
				Basic	2786	2440	1130	1070	2782	2435
				Advanced	672	518	3398	1887	665	562

Table 16 (cont.)

Test Characteristics					Difficulty_Variability Level and Equating Method					
					Moderate/ Reasonable		Easy/ Reasonable		Moderate/ Constrict	
Design	a_ Lvl	Cor_ Lvl	Anc_b change	Prof Level	3 PL Classif	UME_S Classif	3 PL Classif	UME_S Classif	3 PL Classif	UME_S Classif
50_10	High	Mod	-.25	Bel_prof	1535	916	336	190	1581	834
				Basic	2792	1970	2028	905	2759	1297
				Advanced	673	668	2636	2634	660	660
			.00	Bel_prof	1548	1508	345	280	1576	1180
				Basic	2787	2060	2143	1331	2765	1661
				Advanced	665	582	2512	2499	659	659
			+.25	Bel_prof	1517	1500	385	385	1543	1466
				Basic	2805	1854	2176	1444	2813	1942
				Advanced	678	555	2439	1877	644	637
		High	-.25	Bel_prof	1569	1070	281	205	1552	995
				Basic	2734	1692	1968	1017	2752	1386
				Advanced	697	697	2751	2747	696	696
			.00	Bel_prof	1514	1371	327	294	1532	1246
				Basic	2820	2044	2021	1304	2784	1765
				Advanced	666	663	2652	2627	684	682
			+.25	Bel_prof	1583	1537	352	347	1592	1517
				Basic	2736	1978	2056	1584	2714	1876
				Advanced	681	622	2592	2331	694	686

Table 17

Number of Test Takers Correctly Classified for Weighted Mean Equating Method with Stabilization for 50_10 Test Design

Test Characteristics					Difficulty_Variability Level and Equating Method					
					Moderate/ Reasonable		Easy/ Reasonable		Moderate/ Constrict	
Design	a_ Lvl	Cor_ Lvl	Anc_b change	Prof Level	3 PL Classif	WME$_S$ Classif	3 PL Classif	WME$_S$ Classif	3 PL Classif	WME$_S$ Classif
50_10	Mod	Mod	-.25	Bel_prof	1577	553	424	157	1547	720
				Basic	2771	2067	1281	1138	2777	2118
				Advanced	652	627	3295	2847	676	656
			.00	Bel_prof	1572	877	463	214	1595	1036
				Basic	2763	2271	1332	1245	2711	2212
				Advanced	665	613	3205	2538	694	652
			+.25	Bel_prof	1540	1363	521	280	1570	1178
				Basic	2787	2297	1351	1294	2760	2375
				Advanced	673	403	3128	2238	670	568
		High	-.25	Bel_prof	1551	717	417	121	1585	507
				Basic	2774	2267	1054	987	2748	1906
				Advanced	675	627	3529	2819	667	657
			.00	Bel_prof	1560	880	446	194	1551	790
				Basic	2774	2385	1088	1047	2786	2364
				Advanced	666	579	3466	2483	663	610
			+.25	Bel_prof	1542	1198	472	303	1553	1077
				Basic	2786	2440	1130	1070	2782	2435
				Advanced	672	518	3398	1887	665	562

Table 17 (cont.)

Test Characteristics					Difficulty_Variability Level and Equating Method					
					Moderate/ Reasonable		Easy/ Reasonable		Moderate/ Constrict	
			Anc_b		3 PL	WME$_S$	3 PL	WME$_S$	3 PL	WME$_S$
Design	a_ Lvl	Cor_ Lvl	change	Prof Level	Classif	Classif	Classif	Classif	Classif	Classif
50_10	High	Mod	-.25	Bel_prof	1535	916	336	190	1581	834
				Basic	2792	1970	2028	905	2759	1297
				Advanced	673	668	2636	2634	660	660
			.00	Bel_prof	1548	1508	345	280	1576	1180
				Basic	2787	2060	2143	1331	2765	1661
				Advanced	665	582	2512	2499	659	659
			+.25	Bel_prof	1517	1500	385	385	1543	1466
				Basic	2805	1854	2176	1444	2813	1942
				Advanced	678	555	2439	1877	644	637
		High	-.25	Bel_prof	1569	1070	281	205	1552	995
				Basic	2734	1692	1968	1017	2752	1386
				Advanced	697	697	2751	2747	696	696
			.00	Bel_prof	1514	1371	327	294	1532	1246
				Basic	2820	2044	2021	1304	2784	1765
				Advanced	666	663	2652	2627	684	682
			+.25	Bel_prof	1583	1537	352	347	1592	1517
				Basic	2736	1978	2056	1584	2714	1876
				Advanced	681	622	2592	2331	694	686

Table 18

Number of Test Takers Correctly Classified for Unweighted Mean Equating Method for 50_15 Test Design

Test Characteristics					Difficulty_Variability Level and Equating Method					
					Moderate/ Reasonable		Easy/ Reasonable		Moderate/ Constrict	
Design	a_ Lvl	Cor_ Lvl	Anc_b change	Prof Level	3 PL Classif	UME_F Classif	3 PL Classif	UME_F Classif	3 PL Classif	UME_F Classif
50_15	Mod	Mod	-.25	Bel_prof	1562	497	400	119	1569	551
				Basic	2757	1983	1233	1056	2741	2067
				Advanced	681	663	3367	2979	690	667
			.00	Bel_prof	1559	847	441	215	1556	801
				Basic	2750	2336	1378	1247	2751	2395
				Advanced	691	593	3181	2655	693	621
			+.25	Bel_prof	1548	1040	500	306	1569	1127
				Basic	2771	2483	1460	1350	2734	2428
				Advanced	681	549	3040	2328	697	566
		High	-.25	Bel_prof	1543	517	431	127	1585	596
				Basic	2777	1970	1185	976	2730	2020
				Advanced	680	658	3384	3060	685	669
			.00	Bel_prof	1583	781	437	149	1544	798
				Basic	2745	2320	1378	1220	2790	2292
				Advanced	672	598	3185	2708	666	626
			+.25	Bel_prof	1553	1191	526	265	1542	1090
				Basic	2770	2447	1370	1289	2791	2462
				Advanced	677	539	3104	2213	667	532

Table 18 (cont.)

Test Characteristics					Difficulty_Variability Level and Equating Method					
					Moderate/ Reasonable		Easy/ Reasonable		Moderate/ Constrict	
Design	a_ Lvl	Cor_ Lvl	Anc_b change	Prof Level	3 PL Classif	UME$_F$ Classif	3 PL Classif	UME$_F$ Classif	3 PL Classif	UME$_F$ Classif
50_15	High	Mod	-.25	Bel_prof	1536	795	297	168	1553	899
				Basic	2783	1342	1962	816	2771	1506
				Advanced	681	681	2741	2740	676	676
			.00	Bel_prof	1507	1051	349	235	1562	1222
				Basic	2834	1786	2051	1138	2761	2001
				Advanced	659	658	2600	2589	677	676
			+.25	Bel_prof	1562	1355	397	337	1557	1423
				Basic	2753	1988	2182	1501	2748	1949
				Advanced	685	668	2421	2364	695	682
		High	-.25	Bel_prof	1536	820	290	127	1597	873
				Basic	2797	1576	1953	846	2731	1532
				Advanced	667	667	2757	2755	672	672
			.00	Bel_prof	1543	1077	349	202	1537	1219
				Basic	2770	1966	2087	1064	2803	1873
				Advanced	687	681	2564	2562	660	659
			+.25	Bel_prof	1557	1366	408	304	1556	1438
				Basic	2754	2131	2130	1438	2764	2033
				Advanced	689	659	2462	2423	680	670

Table 19

Number of Test Takers Correctly Classified for Weighted Mean Equating Method for 50_15 Test Design

Test Characteristics					Difficulty_Variability Level and Equating Method					
					Moderate/ Reasonable		Easy/ Reasonable		Moderate/ Constrict	
Design	a_ Lvl	Cor_ Lvl	Anc_b change	Prof Level	3 PL Classif	WME$_F$ Classif	3 PL Classif	WME$_F$ Classif	3 PL Classif	WME$_F$ Classif
50_15	Mod	Mod	-.25	Bel_prof	1562	605	400	142	1569	551
				Basic	2757	1977	1233	1055	2741	2067
				Advanced	681	663	3367	2979	690	667
			.00	Bel_prof	1559	717	441	215	1556	801
				Basic	2750	2202	1378	1247	2751	2395
				Advanced	691	628	3181	2655	693	621
			+.25	Bel_prof	1548	1040	500	263	1569	1127
				Basic	2771	2373	1460	1368	2734	2428
				Advanced	681	591	3040	2328	697	566
		High	-.25	Bel_prof	1543	517	431	153	1585	596
				Basic	2777	1970	1185	1049	2730	2020
				Advanced	680	658	3384	2923	685	669
			.00	Bel_prof	1583	781	437	149	1544	798
				Basic	2745	2320	1378	1220	2790	2292
				Advanced	672	598	3185	2708	666	626
			+.25	Bel_prof	1553	1062	526	265	1542	1090
				Basic	2770	2493	1370	1254	2791	2462
				Advanced	677	539	3104	2427	667	532

Table 19 (cont.)

Test Characteristics					Difficulty_Variability Level and Equating Method					
					Moderate/ Reasonable		Easy/ Reasonable		Moderate/ Constrict	
Design	a_ Lvl	Cor_ Lvl	Anc_b change	Prof Level	3 PL Classif	WME$_F$ Classif	3 PL Classif	WME$_F$ Classif	3 PL Classif	WME$_F$ Classif
50_15	High	Mod	-.25	Bel_prof	1536	907	297	168	1553	899
				Basic	2783	1665	1962	816	2771	1506
				Advanced	681	681	2741	2740	676	676
			.00	Bel_prof	1507	1153	349	203	1562	1222
				Basic	2834	1932	2051	995	2761	1876
				Advanced	659	658	2600	2595	677	677
			+.25	Bel_prof	1562	1427	397	242	1557	1423
				Basic	2753	2068	2182	1262	2748	1949
				Advanced	685	650	2421	2406	695	682
		High	-.25	Bel_prof	1536	1027	290	155	1597	873
				Basic	2797	1916	1953	841	2731	1532
				Advanced	667	665	2757	2755	672	672
			.00	Bel_prof	1543	1077	349	202	1537	1219
				Basic	2770	1966	2087	1064	2803	1873
				Advanced	687	681	2564	2562	660	659
			+.25	Bel_prof	1557	1366	408	304	1556	1372
				Basic	2754	2001	2130	1282	2764	2131
				Advanced	689	671	2462	2446	680	670

Table 20

Number of Test Takers Correctly Classified for Anchor Item Calibration Method for 50_15 Test Design

Test Characteristics					Difficulty_Variability Level and Equating Method					
					Moderate/ Reasonable		Easy/ Reasonable		Moderate/ Constrict	
Design	a_ Lvl	Cor_ Lvl	Anc_b change	Prof Level	3 PL Classif	AIC Classif	3 PL Classif	AIC Classif	3 PL Classif	AIC Classif
50_15	Mod	Mod	-.25	Bel_prof	1562	497	400	119	1569	551
				Basic	2757	1983	1233	958	2741	2067
				Advanced	681	663	3367	3111	690	667
			.00	Bel_prof	1559	847	441	215	1556	801
				Basic	2750	2336	1378	1247	2751	2395
				Advanced	691	593	3181	2655	693	621
			+.25	Bel_prof	1548	645	500	306	1569	1127
				Basic	2771	2156	1460	1350	2734	2428
				Advanced	681	651	3040	2328	697	566
		High	-.25	Bel_prof	1543	517	431	107	1585	596
				Basic	2777	1970	1185	977	2730	2020
				Advanced	680	658	3384	3060	685	669
			.00	Bel_prof	1583	781	437	149	1544	798
				Basic	2745	2320	1378	1220	2790	2292
				Advanced	672	598	3185	2708	666	626
			+.25	Bel_prof	1553	1191	526	265	1542	1090
				Basic	2770	2447	1370	1289	2791	2462
				Advanced	677	539	3104	2213	667	532

Table 20 (cont.)

Test Characteristics					Difficulty_Variability Level and Equating Method					
					Moderate/ Reasonable		Easy/ Reasonable		Moderate/ Constrict	
Design	a_ Lvl	Cor_ Lvl	Anc_b change	Prof Level	3 PL Classif	AIC Classif	3 PL Classif	AIC Classif	3 PL Classif	AIC Classif
50_15	High	Mod	-.25	Bel_prof	1536	795	297	168	1553	899
				Basic	2783	1522	1962	816	2771	1506
				Advanced	681	681	2741	2740	676	676
			.00	Bel_prof	1507	1051	349	235	1562	1222
				Basic	2834	1975	2051	1138	2761	2001
				Advanced	659	658	2600	2589	677	676
			+.25	Bel_prof	1562	1355	397	337	1557	1423
				Basic	2753	2144	2182	1354	2748	1949
				Advanced	685	650	2421	2384	695	682
		High	-.25	Bel_prof	1536	820	290	127	1597	873
				Basic	2797	1576	1953	846	2731	1532
				Advanced	667	667	2757	2755	672	672
			.00	Bel_prof	1543	1077	349	202	1537	1219
				Basic	2770	1966	2087	1064	2803	1873
				Advanced	687	681	2564	2562	660	659
			+.25	Bel_prof	1557	1366	408	304	1556	1438
				Basic	2754	2001	2130	1438	2764	2033
				Advanced	689	671	2462	2423	680	670

Table 21

Number of Test Takers Correctly Classified for Unweighted Mean Equating Method with Stabilization for 50_15 Test Design

Test Characteristics					Difficulty_Variability Level and Equating Method					
					Moderate/ Reasonable		Easy/ Reasonable		Moderate/ Constrict	
Design	a_ Lvl	Cor_ Lvl	Anc_b change	Prof Level	3 PL Classif	UME$_S$ Classif	3 PL Classif	UME$_S$ Classif	3 PL Classif	UME$_S$ Classif
50_15	Mod	Mod	-.25	Bel_prof	1562	321	400	96	1569	551
				Basic	2757	1353	1233	868	2741	2067
				Advanced	681	680	3367	3197	690	667
			.00	Bel_prof	1559	393	441	143	1556	801
				Basic	2750	1695	1378	1073	2751	2260
				Advanced	691	680	3181	2929	693	653
			+.25	Bel_prof	1548	645	500	426	1569	1127
				Basic	2771	2156	1460	1311	2734	2428
				Advanced	681	651	3040	1570	697	566
		High	-.25	Bel_prof	1543	517	431	127	1585	596
				Basic	2777	1970	1185	976	2730	2020
				Advanced	680	658	3384	3060	685	669
			.00	Bel_prof	1583	781	437	149	1544	798
				Basic	2745	2320	1378	1220	2790	2292
				Advanced	672	598	3185	2708	666	626
			+.25	Bel_prof	1553	1191	526	265	1542	1090
				Basic	2770	2447	1370	1289	2791	2462
				Advanced	677	539	3104	2213	667	532

Table 21 (cont.)

Test Characteristics					Difficulty_Variability Level and Equating Method					
					Moderate/ Reasonable		Easy/ Reasonable		Moderate/ Constrict	
Design	a_ Lvl	Cor_ Lvl	Anc_b change	Prof Level	3 PL Classif	UME_S Classif	3 PL Classif	UME_S Classif	3 PL Classif	UME_S Classif
50_15	High	Mod	-.25	Bel_prof	1536	385	297	262	1553	781
				Basic	2783	764	1962	1436	2771	1396
				Advanced	681	681	2741	2656	676	676
			.00	Bel_prof	1507	625	349	340	1562	1105
				Basic	2834	1321	2051	1612	2761	1787
				Advanced	659	659	2600	2339	677	677
			+.25	Bel_prof	1562	935	397	279	1557	1361
				Basic	2753	1712	2182	1254	2748	1941
				Advanced	685	685	2421	2406	695	688
		High	-.25	Bel_prof	1536	820	290	127	1597	873
				Basic	2797	1576	1953	846	2731	1532
				Advanced	667	667	2757	2755	672	672
			.00	Bel_prof	1543	1077	349	228	1537	1219
				Basic	2770	1966	2087	1247	2803	1873
				Advanced	687	681	2564	2547	660	659
			+.25	Bel_prof	1557	1366	408	47	1556	1438
				Basic	2754	2001	2130	233	2764	2033
				Advanced	689	671	2462	2462	680	670

Table 22

Number of Test Takers Correctly Classified for Weighted Mean Equating Method with Stabilization for 50_15 Test Design

Test Characteristics					Difficulty_Variability Level and Equating Method					
					Moderate/ Reasonable		Easy/ Reasonable		Moderate/ Constrict	
Design	a_ Lvl	Cor_ Lvl	Anc_b change	Prof Level	3 PL Classif	WME$_S$ Classif	3 PL Classif	WME$_S$ Classif	3 PL Classif	WME$_S$ Classif
50_15	Mod	Mod	-.25	Bel_prof	1562	409	400	311	1569	551
				Basic	2757	1786	1233	1119	2741	2067
				Advanced	681	671	3367	2161	690	667
			.00	Bel_prof	1559	393	441	342	1556	801
				Basic	2750	1695	1378	1257	2751	2260
				Advanced	691	680	3181	1976	693	653
			+.25	Bel_prof	1548	645	500	182	1569	1127
				Basic	2771	2156	1460	1226	2734	2428
				Advanced	681	651	3040	2680	697	566
		High	-.25	Bel_prof	1543	517	431	127	1585	596
				Basic	2777	1970	1185	976	2730	2020
				Advanced	680	658	3384	3060	685	669
			.00	Bel_prof	1583	781	437	149	1544	798
				Basic	2745	2320	1378	1220	2790	2292
				Advanced	672	598	3185	2708	666	626
			+.25	Bel_prof	1553	1191	526	265	1542	1090
				Basic	2770	2447	1370	1289	2791	2462
				Advanced	677	539	3104	2213	667	532

Table 22 (cont.)

Test Characteristics					Difficulty_Variability Level and Equating Method					
					Moderate/ Reasonable		Easy/ Reasonable		Moderate/ Constrict	
Design	a_ Lvl	Cor_ Lvl	Anc_b change	Prof Level	3 PL Classif	WME$_S$ Classif	3 PL Classif	WME$_S$ Classif	3 PL Classif	WME$_S$ Classif
50_15	High	Mod	-.25	Bel_prof	1536	1395	297	117	1553	781
				Basic	2783	2059	1962	541	2771	1396
				Advanced	681	667	2741	2741	676	676
			.00	Bel_prof	1507	1473	349	148	1562	1105
				Basic	2834	2062	2051	852	2761	1787
				Advanced	659	588	2600	2600	677	677
			+.25	Bel_prof	1562	1164	397	279	1557	1361
				Basic	2753	1980	2182	1254	2748	1941
				Advanced	685	680	2421	2406	695	688
		High	-.25	Bel_prof	1536	820	290	127	1597	873
				Basic	2797	1576	1953	846	2731	1532
				Advanced	667	667	2757	2755	672	672
			.00	Bel_prof	1543	1077	349	202	1537	1219
				Basic	2770	1966	2087	1255	2803	1873
				Advanced	687	681	2564	2547	660	659
			+.25	Bel_prof	1557	1366	408	330	1556	1438
				Basic	2754	2001	2130	1409	2764	2033
				Advanced	689	671	2462	2423	680	670

Table 23

Number of Test Takers Correctly Classified for Unweighted Mean Equating Method for 100_20 Test Design

Test Characteristics					Difficulty_Variability Level and Equating Method					
					Moderate/ Reasonable		Easy/ Reasonable		Moderate/ Constrict	
Design	a_ Lvl	Cor_ Lvl	Anc_b change	Prof Level	3 PL Classif	UME$_F$ Classif	3 PL Classif	UME$_F$ Classif	3 PL Classif	UME$_F$ Classif
100_20	Mod	Mod	-.25	Bel_prof	1555	199	391	61	1549	392
				Basic	2737	1746	1121	803	2741	1765
				Advanced	708	707	3488	3392	710	709
			.00	Bel_prof	1568	477	420	86	1547	600
				Basic	2745	2350	1210	1053	2763	2184
				Advanced	687	670	3370	3088	690	678
			+.25	Bel_prof	1565	642	444	163	1562	895
				Basic	2757	2463	1300	1244	2754	2441
				Advanced	678	620	3256	2678	684	650
		High	-.25	Bel_prof	1535	392	327	62	1558	370
				Basic	2791	2513	1241	987	2768	1816
				Advanced	674	572	3432	3283	674	674
			.00	Bel_prof	1552	554	386	114	1567	642
				Basic	2785	2598	1262	1167	2739	2214
				Advanced	663	540	3352	3023	694	685
			+.25	Bel_prof	1552	684	429	214	1563	1004
				Basic	2760	2650	1313	1295	2779	2519
				Advanced	688	518	3258	2531	658	587

Table 23 (cont.)

Test Characteristics					Difficulty_Variability Level and Equating Method					
					Moderate/ Reasonable		Easy/ Reasonable		Moderate/ Constrict	
Design	a_ Lvl	Cor_ Lvl	Anc_b change	Prof Level	3 PL Classif	UME_F Classif	3 PL Classif	UME_F Classif	3 PL Classif	UME_F Classif
100_20	High	Mod	-.25	Bel_prof	1558	537	266	96	1569	721
				Basic	2774	1063	2026	525	2743	1275
				Advanced	668	668	2708	2708	688	688
			.00	Bel_prof	1545	1063	280	148	1538	1114
				Basic	2798	1818	2104	837	2804	1743
				Advanced	657	657	2616	2616	658	658
			+.25	Bel_prof	1552	1164	313	247	1564	1431
				Basic	2773	2487	2167	1248	2777	1997
				Advanced	675	642	2520	2519	659	658
		High	-.25	Bel_prof	1565	859	262	97	1573	746
				Basic	2757	1603	1961	472	2736	1258
				Advanced	678	678	2777	2777	691	691
			.00	Bel_prof	1576	1198	290	166	1607	1116
				Basic	2738	1982	2061	832	2724	1717
				Advanced	686	685	2649	2649	669	669
			+.25	Bel_prof	1533	788	302	259	1564	1416
				Basic	2785	2135	2105	1267	2763	2007
				Advanced	682	678	2593	2593	673	673

Table 24

Number of Test Takers Correctly Classified for Weighted Mean Equating Method for 100_20 Test Design

Test Characteristics					Difficulty_Variability Level and Equating Method					
					Moderate/ Reasonable		Easy/ Reasonable		Moderate/ Constrict	
Design	a_ Lvl	Cor_ Lvl	Anc_b change	Prof Level	3 PL Classif	WME$_F$ Classif	3 PL Classif	WME$_F$ Classif	3 PL Classif	WME$_F$ Classif
100_20	Mod	Mod	-.25	Bel_prof	1555	266	391	84	1549	392
				Basic	2737	1974	1121	924	2741	1765
				Advanced	708	704	3488	3292	710	709
			.00	Bel_prof	1568	543	420	86	1547	658
				Basic	2745	2349	1210	1053	2763	2279
				Advanced	687	670	3370	3088	690	674
			+.25	Bel_prof	1565	810	444	143	1562	895
				Basic	2757	2508	1300	1244	2754	2441
				Advanced	678	599	3256	2678	684	650
		High	-.25	Bel_prof	1535	392	327	85	1558	427
				Basic	2791	2513	1241	1051	2768	1816
				Advanced	674	572	3432	3222	674	674
			.00	Bel_prof	1552	554	386	114	1567	642
				Basic	2785	2598	1262	1167	2739	2214
				Advanced	663	540	3352	3023	694	685
			+.25	Bel_prof	1552	684	429	198	1563	935
				Basic	2760	2650	1313	1290	2779	2528
				Advanced	688	518	3258	2636	658	587

Table 24 (cont.)

Test Characteristics					Difficulty_Variability Level and Equating Method					
					Moderate/ Reasonable		Easy/ Reasonable		Moderate/ Constrict	
Design	a_ Lvl	Cor_ Lvl	Anc_b change	Prof Level	3 PL Classif	WME_F Classif	3 PL Classif	WME_F Classif	3 PL Classif	WME_F Classif
100_20	High	Mod	-.25	Bel_prof	1558	762	266	114	1569	721
				Basic	2774	1496	2026	577	2743	1340
				Advanced	668	668	2708	2708	688	688
			.00	Bel_prof	1545	1132	280	156	1538	1114
				Basic	2798	1978	2104	837	2804	1743
				Advanced	657	656	2616	2616	658	658
			+.25	Bel_prof	1552	942	313	247	1564	1345
				Basic	2773	2280	2167	1168	2777	1973
				Advanced	675	663	2520	2519	659	659
		High	-.25	Bel_prof	1565	997	262	124	1573	798
				Basic	2757	1766	1961	651	2736	1323
				Advanced	678	678	2777	2777	691	691
			.00	Bel_prof	1576	1198	290	166	1607	1116
				Basic	2738	1982	2061	832	2724	1717
				Advanced	686	685	2649	2649	669	669
			+.25	Bel_prof	1533	712	302	245	1564	1382
				Basic	2785	2050	2105	1186	2763	1969
				Advanced	682	679	2593	2593	673	673

Table 25

Number of Test Takers Correctly Classified for Anchor Item Calibration Method for 100_20 Test Design

Test Characteristics					Difficulty_Variability Level and Equating Method					
					Moderate/ Reasonable		Easy/ Reasonable		Moderate/ Constrict	
Design	a_ Lvl	Cor_ Lvl	Anc_b change	Prof Level	3 PL Classif	AIC Classif	3 PL Classif	AIC Classif	3 PL Classif	AIC Classif
100_20	Mod	Mod	-.25	Bel_prof	1555	199	391	61	1549	392
				Basic	2737	1746	1121	803	2741	1765
				Advanced	708	707	3488	3392	710	709
			.00	Bel_prof	1568	543	420	86	1547	600
				Basic	2745	2349	1210	1053	2763	2184
				Advanced	687	670	3370	3088	690	678
			+.25	Bel_prof	1565	642	444	163	1562	895
				Basic	2757	2463	1300	1244	2754	2441
				Advanced	678	620	3256	2678	684	650
		High	-.25	Bel_prof	1535	392	327	62	1558	370
				Basic	2791	2513	1241	987	2768	1816
				Advanced	674	572	3432	3283	674	674
			.00	Bel_prof	1552	554	386	114	1567	642
				Basic	2785	2598	1262	1203	2739	2124
				Advanced	663	540	3352	2936	694	688
			+.25	Bel_prof	1552	684	429	214	1563	1004
				Basic	2760	2650	1313	1295	2779	2519
				Advanced	688	518	3258	2531	658	587

Table 25 (cont.)

Test Characteristics					Difficulty_Variability Level and Equating Method					
					Moderate/ Reasonable		Easy/ Reasonable		Moderate/ Constrict	
Design	a_ Lvl	Cor_ Lvl	Anc_b change	Prof Level	3 PL Classif	AIC Classif	3 PL Classif	AIC Classif	3 PL Classif	AIC Classif
100_20	High	Mod	-.25	Bel_prof	1558	537	266	168	1569	721
				Basic	2774	1135	2026	934	2743	1275
				Advanced	668	668	2708	2708	688	688
			.00	Bel_prof	1545	1063	280	167	1538	1114
				Basic	2798	1905	2104	930	2804	1743
				Advanced	657	657	2616	2616	658	658
			+.25	Bel_prof	1552	1164	313	192	1564	1393
				Basic	2773	2414	2167	885	2777	2028
				Advanced	675	651	2520	2520	659	658
		High	-.25	Bel_prof	1565	859	262	178	1573	746
				Basic	2757	1603	1961	870	2736	1323
				Advanced	678	678	2777	2777	691	691
			.00	Bel_prof	1576	1198	290	183	1607	1116
				Basic	2738	1982	2061	831	2724	1717
				Advanced	686	685	2649	2649	669	669
			+.25	Bel_prof	1533	861	302	170	1564	1382
				Basic	2785	2132	2105	845	2763	2038
				Advanced	682	678	2593	2593	673	673

Table 26

Number of Test Takers Correctly Classified for Unweighted Mean Equating Method with Stabilization for 100_20 Test Design

Test Characteristics					Difficulty_Variability Level and Equating Method					
					Moderate/ Reasonable		Easy/ Reasonable		Moderate/ Constrict	
Design	a_ Lvl	Cor_ Lvl	Anc_b change	Prof Level	3 PL Classif	UME$_S$ Classif	3 PL Classif	UME$_S$ Classif	3 PL Classif	UME$_S$ Classif
100_20	Mod	Mod	-.25	Bel_prof	1555	490	391	26	1549	392
				Basic	2737	2276	1121	676	2741	1874
				Advanced	708	685	3488	3439	710	709
			.00	Bel_prof	1568	410	420	71	1547	600
				Basic	2745	2264	1210	1006	2763	2184
				Advanced	687	676	3370	3146	690	678
			+.25	Bel_prof	1565	569	444	163	1562	813
				Basic	2757	2374	1300	1262	2754	2444
				Advanced	678	637	3256	2571	684	650
		High	-.25	Bel_prof	1535	392	327	62	1558	370
				Basic	2791	2513	1241	987	2768	1816
				Advanced	674	572	3432	3283	674	674
			.00	Bel_prof	1552	554	386	114	1567	642
				Basic	2785	2598	1262	1203	2739	2124
				Advanced	663	540	3352	2936	694	688
			+.25	Bel_prof	1552	684	429	214	1563	1004
				Basic	2760	2650	1313	1295	2779	2519
				Advanced	688	518	3258	2531	658	587

Table 26 (cont.)

Test Characteristics					Difficulty_Variability Level and Equating Method					
					Moderate/ Reasonable		Easy/ Reasonable		Moderate/ Constrict	
Design	a_ Lvl	Cor_ Lvl	Anc_b change	Prof Level	3 PL Classif	UME$_S$ Classif	3 PL Classif	UME$_S$ Classif	3 PL Classif	UME$_S$ Classif
100_20	High	Mod	-.25	Bel_prof	1558	877	266	77	1569	624
				Basic	2774	1680	2026	426	2743	1147
				Advanced	668	668	2708	2708	688	688
			.00	Bel_prof	1545	1456	280	156	1538	1049
				Basic	2798	2187	2104	837	2804	1683
				Advanced	657	645	2616	2616	658	658
			+.25	Bel_prof	1552	319	313	247	1564	1345
				Basic	2773	1368	2167	1168	2777	1973
				Advanced	675	675	2520	2519	659	659
		High	-.25	Bel_prof	1565	859	262	97	1573	746
				Basic	2757	1603	1961	472	2736	1323
				Advanced	678	678	2777	2777	691	691
			.00	Bel_prof	1576	1198	290	166	1607	1116
				Basic	2738	1982	2061	832	2724	1717
				Advanced	686	685	2649	2649	669	669
			+.25	Bel_prof	1533	335	302	259	1564	1138
				Basic	2785	1429	2105	1267	2763	1839
				Advanced	682	682	2593	2593	673	673

Table 27

Number of Test Takers Correctly Classified for Weighted Mean Equating Method with Stabilization for 100_20 Test Design

Test Characteristics					Difficulty_Variability Level and Equating Method					
					Moderate/ Reasonable		Easy/ Reasonable		Moderate/ Constrict	
Design	a_ Lvl	Cor_ Lvl	Anc_b change	Prof Level	3 PL Classif	WME$_S$ Classif	3 PL Classif	WME$_S$ Classif	3 PL Classif	WME$_S$ Classif
100_20	Mod	Mod	-.25	Bel_prof	1555	559	391	106	1549	392
				Basic	2737	2355	1121	1017	2741	1874
				Advanced	708	671	3488	3124	710	709
			.00	Bel_prof	1568	884	420	71	1547	600
				Basic	2745	2606	1210	1006	2763	2184
				Advanced	687	583	3370	3146	690	678
			+.25	Bel_prof	1565	899	444	101	1562	813
				Basic	2757	2571	1300	1198	2754	2376
				Advanced	678	569	3256	2901	684	662
		High	-.25	Bel_prof	1535	392	327	62	1558	370
				Basic	2791	2513	1241	987	2768	1816
				Advanced	674	572	3432	3283	674	674
			.00	Bel_prof	1552	554	386	114	1567	642
				Basic	2785	2598	1262	1203	2739	2124
				Advanced	663	540	3352	2936	694	688
			+.25	Bel_prof	1552	684	429	214	1563	1004
				Basic	2760	2650	1313	1295	2779	2519
				Advanced	688	518	3258	2531	658	587

Table 27 (cont.)

Test Characteristics					Difficulty_Variability Level and Equating Method					
					Moderate/ Reasonable		Easy/ Reasonable		Moderate/ Constrict	
			Anc_b		3 PL	WME$_S$	3 PL	WME$_S$	3 PL	WME$_S$
Design	a_ Lvl	Cor_ Lvl	change	Prof Level	Classif	Classif	Classif	Classif	Classif	Classif
100_20	High	Mod	-.25	Bel_prof	1558	1027	266	96	1569	665
				Basic	2774	1784	2026	467	2743	1215
				Advanced	668	668	2708	2708	688	688
			.00	Bel_prof	1545	1488	280	148	1538	1114
				Basic	2798	2234	2104	837	2804	1743
				Advanced	657	640	2616	2616	658	658
			+.25	Bel_prof	1552	319	313	247	1564	1345
				Basic	2773	1368	2167	1168	2777	1973
				Advanced	675	675	2520	2519	659	659
		High	-.25	Bel_prof	1565	859	262	97	1573	746
				Basic	2757	1603	1961	472	2736	1323
				Advanced	678	678	2777	2777	691	691
			.00	Bel_prof	1576	1198	290	166	1607	1116
				Basic	2738	1982	2061	832	2724	1717
				Advanced	686	685	2649	2649	669	669
			+.25	Bel_prof	1533	335	302	259	1564	1138
				Basic	2785	1429	2105	1267	2763	1839
				Advanced	682	682	2593	2593	673	673

Table 28

Number of Test Takers Correctly Classified for Unweighted Mean Equating Method for 100_30 Test Design

Test Characteristics					Difficulty_Variability Level and Equating Method					
					Moderate/ Reasonable		Easy/ Reasonable		Moderate/ Constrict	
Design	a_ Lvl	Cor_ Lvl	Anc_b change	Prof Level	3 PL Classif	UME$_F$ Classif	3 PL Classif	UME$_F$ Classif	3 PL Classif	UME$_F$ Classif
100_30	Mod	Mod	-.25	Bel_prof	1557	322	437	74	1539	515
				Basic	2791	1911	1302	951	2791	2097
				Advanced	652	643	3261	3177	670	665
			.00	Bel_prof	1512	453	491	126	1588	1096
				Basic	2811	2434	1383	1224	2731	2548
				Advanced	677	630	3126	2877	681	563
			+.25	Bel_prof	1556	720	545	223	1554	1077
				Basic	2746	2595	1447	1400	2758	2607
				Advanced	698	584	3008	2487	688	566
		High	-.25	Bel_prof	1548	589	422	74	1554	476
				Basic	2754	2242	1284	1015	2786	1920
				Advanced	698	677	3294	3150	660	660
			.00	Bel_prof	1530	708	479	154	1566	709
				Basic	2808	2465	1335	1242	2751	2321
				Advanced	662	618	3186	2842	683	670
			+.25	Bel_prof	1585	964	517	232	1543	1107
				Basic	2729	2565	1494	1455	2783	2595
				Advanced	686	587	2989	2351	674	587

Table 28 (cont.)

Test Characteristics					Difficulty_Variability Level and Equating Method					
					Moderate/ Reasonable		Easy/ Reasonable		Moderate/ Constrict	
Design	a_ Lvl	Cor_ Lvl	Anc_b change	Prof Level	3 PL Classif	UME$_F$ Classif	3 PL Classif	UME$_F$ Classif	3 PL Classif	UME$_F$ Classif
100_30	High	Mod	-.25	Bel_prof	1561	505	312	163	1566	940
				Basic	2743	1123	2084	796	2763	1501
				Advanced	696	696	2604	2604	671	671
			.00	Bel_prof	1576	917	337	255	1566	1307
				Basic	2716	1653	2225	1191	2754	1899
				Advanced	708	707	2438	2438	680	680
			+.25	Bel_prof	1537	1196	402	374	1541	1513
				Basic	2801	2068	2276	1552	2812	1991
				Advanced	662	662	2322	2316	647	644
		High	-.25	Bel_prof	1532	474	324	151	1587	906
				Basic	2811	1046	2088	691	2750	1422
				Advanced	657	657	2588	2588	663	663
			.00	Bel_prof	1584	756	350	249	1583	1296
				Basic	2693	1478	2198	1144	2738	1828
				Advanced	723	723	2452	2452	679	679
			+.25	Bel_prof	1563	1089	386	333	1571	1517
				Basic	2765	1968	2307	1567	2746	1959
				Advanced	672	672	2307	2304	683	681

Table 29

Number of Test Takers Correctly Classified for Weighted Mean Equating Method for 100_30 Test Design

Test Characteristics					Difficulty_Variability Level and Equating Method					
					Moderate/ Reasonable		Easy/ Reasonable		Moderate/ Constrict	
Design	a_ Lvl	Cor_ Lvl	Anc_b change	Prof Level	3 PL Classif	WME$_F$ Classif	3 PL Classif	WME$_F$ Classif	3 PL Classif	WME$_F$ Classif
100_30	Mod	Mod	-.25	Bel_prof	1557	275	437	74	1539	515
				Basic	2791	1911	1302	951	2791	2097
				Advanced	652	643	3261	3177	670	665
			.00	Bel_prof	1512	357	491	109	1588	1096
				Basic	2811	2136	1383	1165	2731	2548
				Advanced	677	662	3126	2952	681	563
			+.25	Bel_prof	1556	459	545	157	1554	1077
				Basic	2746	2399	1447	1342	2758	2607
				Advanced	698	654	3008	2700	688	566
		High	-.25	Bel_prof	1548	519	422	88	1554	519
				Basic	2754	2244	1284	1143	2786	1920
				Advanced	698	677	3294	3035	660	660
			.00	Bel_prof	1530	708	479	154	1566	709
				Basic	2808	2465	1335	1267	2751	2321
				Advanced	662	618	3186	2753	683	670
			+.25	Bel_prof	1585	870	517	186	1543	1107
				Basic	2729	2582	1494	1441	2783	2595
				Advanced	686	587	2989	2454	674	587

Table 29 (cont.)

Test Characteristics					Difficulty_Variability Level and Equating Method					
					Moderate/ Reasonable		Easy/ Reasonable		Moderate/ Constrict	
Design	a_ Lvl	Cor_ Lvl	Anc_b change	Prof Level	3 PL Classif	WME$_F$ Classif	3 PL Classif	WME$_F$ Classif	3 PL Classif	WME$_F$ Classif
100_30	High	Mod	-.25	Bel_prof	1561	550	312	163	1566	1017
				Basic	2743	1295	2084	736	2763	1644
				Advanced	696	696	2604	2604	671	671
			.00	Bel_prof	1576	773	337	198	1566	1362
				Basic	2716	1580	2225	1036	2754	1875
				Advanced	708	708	2438	2438	680	680
			+.25	Bel_prof	1537	887	402	260	1541	1493
				Basic	2801	1826	2276	1180	2812	1972
				Advanced	662	662	2322	2322	647	646
		High	-.25	Bel_prof	1532	634	324	182	1587	975
				Basic	2811	1434	2088	919	2750	1497
				Advanced	657	657	2588	2588	663	663
			.00	Bel_prof	1584	696	350	276	1583	1296
				Basic	2693	1478	2198	1234	2738	1828
				Advanced	723	723	2452	2452	679	679
			+.25	Bel_prof	1563	930	386	286	1571	1492
				Basic	2765	1808	2307	1397	2746	2013
				Advanced	672	672	2307	2307	683	681

Table 30

Number of Test Takers Correctly Classified for Anchor Item Calibration Method for 100_30 Test Design

Test Characteristics					Difficulty_Variability Level and Equating Method					
					Moderate/ Reasonable		Easy/ Reasonable		Moderate/ Constrict	
Design	a_ Lvl	Cor_ Lvl	Anc_b change	Prof Level	3 PL Classif	AIC Classif	3 PL Classif	AIC Classif	3 PL Classif	AIC Classif
100_30	Mod	Mod	-.25	Bel_prof	1557	322	437	88	1539	515
				Basic	2791	1911	1302	874	2791	2097
				Advanced	652	643	3261	3206	670	665
			.00	Bel_prof	1512	453	491	126	1588	1096
				Basic	2811	2331	1383	1224	2731	2497
				Advanced	677	649	3126	2877	681	589
			+.25	Bel_prof	1556	720	545	223	1554	1077
				Basic	2746	2595	1447	1400	2758	2607
				Advanced	698	584	3008	2487	688	566
		High	-.25	Bel_prof	1548	727	422	79	1554	476
				Basic	2754	2413	1284	1080	2786	1920
				Advanced	698	656	3294	3105	660	660
			.00	Bel_prof	1530	708	479	154	1566	783
				Basic	2808	2465	1335	1267	2751	2318
				Advanced	662	618	3186	2753	683	670
			+.25	Bel_prof	1585	964	517	232	1543	1107
				Basic	2729	2565	1494	1455	2783	2548
				Advanced	686	587	2989	2351	674	613

Table 30 (cont.)

Test Characteristics					Difficulty_Variability Level and Equating Method					
					Moderate/ Reasonable		Easy/ Reasonable		Moderate/ Constrict	
Design	a_ Lvl	Cor_ Lvl	Anc_b change	Prof Level	3 PL Classif	AIC Classif	3 PL Classif	AIC Classif	3 PL Classif	AIC Classif
100_30	High	Mod	-.25	Bel_prof	1561	550	312	163	1566	940
				Basic	2743	1123	2084	736	2763	1501
				Advanced	696	696	2604	2604	671	671
			.00	Bel_prof	1576	917	337	255	1566	1307
				Basic	2716	1576	2225	1110	2754	1899
				Advanced	708	708	2438	2438	680	680
			+.25	Bel_prof	1537	1196	402	361	1541	1513
				Basic	2801	1982	2276	1485	2812	1991
				Advanced	662	662	2322	2317	647	644
		High	-.25	Bel_prof	1532	474	324	151	1587	906
				Basic	2811	1046	2088	691	2750	1422
				Advanced	657	657	2588	2588	663	663
			.00	Bel_prof	1584	756	350	230	1583	1296
				Basic	2693	1478	2198	1150	2738	1828
				Advanced	723	723	2452	2452	679	679
			+.25	Bel_prof	1563	1089	386	333	1571	1517
				Basic	2765	1968	2307	1567	2746	1959
				Advanced	672	672	2307	2304	683	681

Table 31

Number of Test Takers Correctly Classified for Unweighted Mean Equating Method with Stabilization for 100_30 Test Design

Test Characteristics					Difficulty_Variability Level and Equating Method					
					Moderate/ Reasonable		Easy/ Reasonable		Moderate/ Constrict	
Design	a_ Lvl	Cor_ Lvl	Anc_b change	Prof Level	3 PL Classif	UME$_S$ Classif	3 PL Classif	UME$_S$ Classif	3 PL Classif	UME$_S$ Classif
100_30	Mod	Mod	-.25	Bel_prof	1557	236	437	164	1539	515
				Basic	2791	1787	1302	1227	2791	2097
				Advanced	652	650	3261	2864	670	665
			.00	Bel_prof	1512	411	491	184	1588	866
				Basic	2811	2331	1383	1338	2731	2483
				Advanced	677	649	3126	2611	681	605
			+.25	Bel_prof	1556	887	545	279	1554	913
				Basic	2746	2644	1447	1411	2758	2532
				Advanced	698	500	3008	2377	688	627
		High	-.25	Bel_prof	1548	401	422	79	1554	476
				Basic	2754	1973	1284	1080	2786	1920
				Advanced	698	690	3294	3105	660	660
			.00	Bel_prof	1530	708	479	154	1566	783
				Basic	2808	2465	1335	1267	2751	2318
				Advanced	662	618	3186	2753	683	670
			+.25	Bel_prof	1585	964	517	232	1543	1107
				Basic	2729	2565	1494	1455	2783	2548
				Advanced	686	587	2989	2351	674	613

Table 31 (cont.)

Test Characteristics					Difficulty_Variability Level and Equating Method					
					Moderate/ Reasonable		Easy/ Reasonable		Moderate/ Constrict	
Design	a_ Lvl	Cor_ Lvl	Anc_b change	Prof Level	3 PL Classif	UME$_S$ Classif	3 PL Classif	UME$_S$ Classif	3 PL Classif	UME$_S$ Classif
100_30	High	Mod	-.25	Bel_prof	1561	405	312	163	1566	726
				Basic	2743	972	2084	736	2763	1330
				Advanced	696	696	2604	2604	671	671
			.00	Bel_prof	1576	842	337	233	1566	1054
				Basic	2716	1580	2225	1115	2754	1638
				Advanced	708	708	2438	2438	680	680
			+.25	Bel_prof	1537	1376	402	382	1541	1201
				Basic	2801	2249	2276	1642	2812	1901
				Advanced	662	661	2322	2303	647	647
		High	-.25	Bel_prof	1532	474	324	151	1587	906
				Basic	2811	1046	2088	762	2750	1422
				Advanced	657	657	2588	2588	663	663
			.00	Bel_prof	1584	756	350	249	1583	1296
				Basic	2693	1478	2198	1144	2738	1828
				Advanced	723	723	2452	2452	679	679
			+.25	Bel_prof	1563	1089	386	333	1571	1200
				Basic	2765	1968	2307	1567	2746	1797
				Advanced	672	672	2307	2304	683	683

Table 32

Number of Test Takers Correctly Classified for Weighted Mean Equating Method with Stabilization for 100_30 Test Design

Test Characteristics					Difficulty_Variability Level and Equating Method					
					Moderate/ Reasonable		Easy/ Reasonable		Moderate/ Constrict	
Design	a_ Lvl	Cor_ Lvl	Anc_b change	Prof Level	3 PL Classif	WME$_S$ Classif	3 PL Classif	WME$_S$ Classif	3 PL Classif	WME$_S$ Classif
100_30	Mod	Mod	-.25	Bel_prof	1557	275	437	164	1539	515
				Basic	2791	1787	1302	1227	2791	2097
				Advanced	652	650	3261	2864	670	665
			.00	Bel_prof	1512	313	491	184	1588	866
				Basic	2811	2137	1383	1338	2731	2483
				Advanced	677	662	3126	2611	681	605
			+.25	Bel_prof	1556	292	545	279	1554	913
				Basic	2746	2107	1447	1411	2758	2532
				Advanced	698	688	3008	2377	688	627
		High	-.25	Bel_prof	1548	401	422	79	1554	476
				Basic	2754	1973	1284	1080	2786	1920
				Advanced	698	690	3294	3105	660	660
			.00	Bel_prof	1530	708	479	154	1566	783
				Basic	2808	2465	1335	1267	2751	2318
				Advanced	662	618	3186	2753	683	670
			+.25	Bel_prof	1585	964	517	232	1543	1107
				Basic	2729	2565	1494	1455	2783	2548
				Advanced	686	587	2989	2351	674	613

Table 32 (cont.)

Test Characteristics					Difficulty_Variability Level and Equating Method					
					Moderate/ Reasonable		Easy/ Reasonable		Moderate/ Constrict	
Design	a_ Lvl	Cor_ Lvl	Anc_b change	Prof Level	3 PL Classif	WME$_S$ Classif	3 PL Classif	WME$_S$ Classif	3 PL Classif	WME$_S$ Classif
100_30	High	Mod	-.25	Bel_prof	1561	1021	312	212	1566	881
				Basic	2743	1794	2084	939	2763	1445
				Advanced	696	696	2604	2604	671	671
			.00	Bel_prof	1576	987	337	304	1566	1002
				Basic	2716	1826	2225	1514	2754	1567
				Advanced	708	707	2438	2438	680	680
			+.25	Bel_prof	1537	627	402	334	1541	1201
				Basic	2801	1413	2276	1412	2812	1901
				Advanced	662	662	2322	2319	647	647
		High	-.25	Bel_prof	1532	474	324	151	1587	906
				Basic	2811	1142	2088	762	2750	1422
				Advanced	657	657	2588	2588	663	663
			.00	Bel_prof	1584	756	350	249	1583	1296
				Basic	2693	1478	2198	1144	2738	1828
				Advanced	723	723	2452	2452	679	679
			+.25	Bel_prof	1563	1089	386	333	1571	1200
				Basic	2765	1968	2307	1567	2746	1797
				Advanced	672	672	2307	2304	683	683

APPENDIX C

MODIFIED BOX PLOTS SHOWING BIAS BY EQUATING METHOD

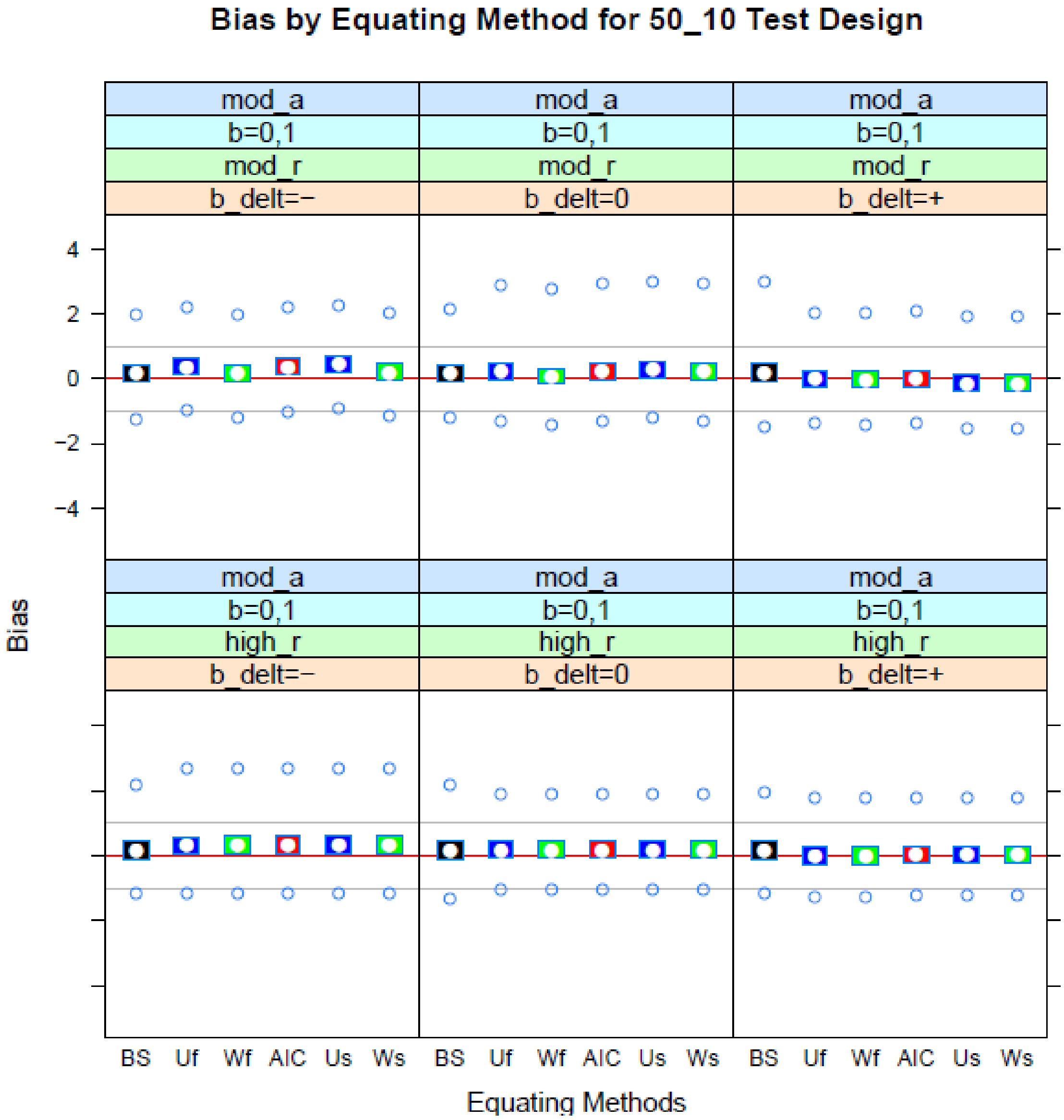


Figure 17. Modified box plot showing bias by equating method. (B = Base form, U = unweighted mean equating, W = weighted mean equating, A = Anchor Item Calibration, U_s = stabilized unweighted mean equating with stabilization and W_s = stabilized weighted mean equating with stabilization)

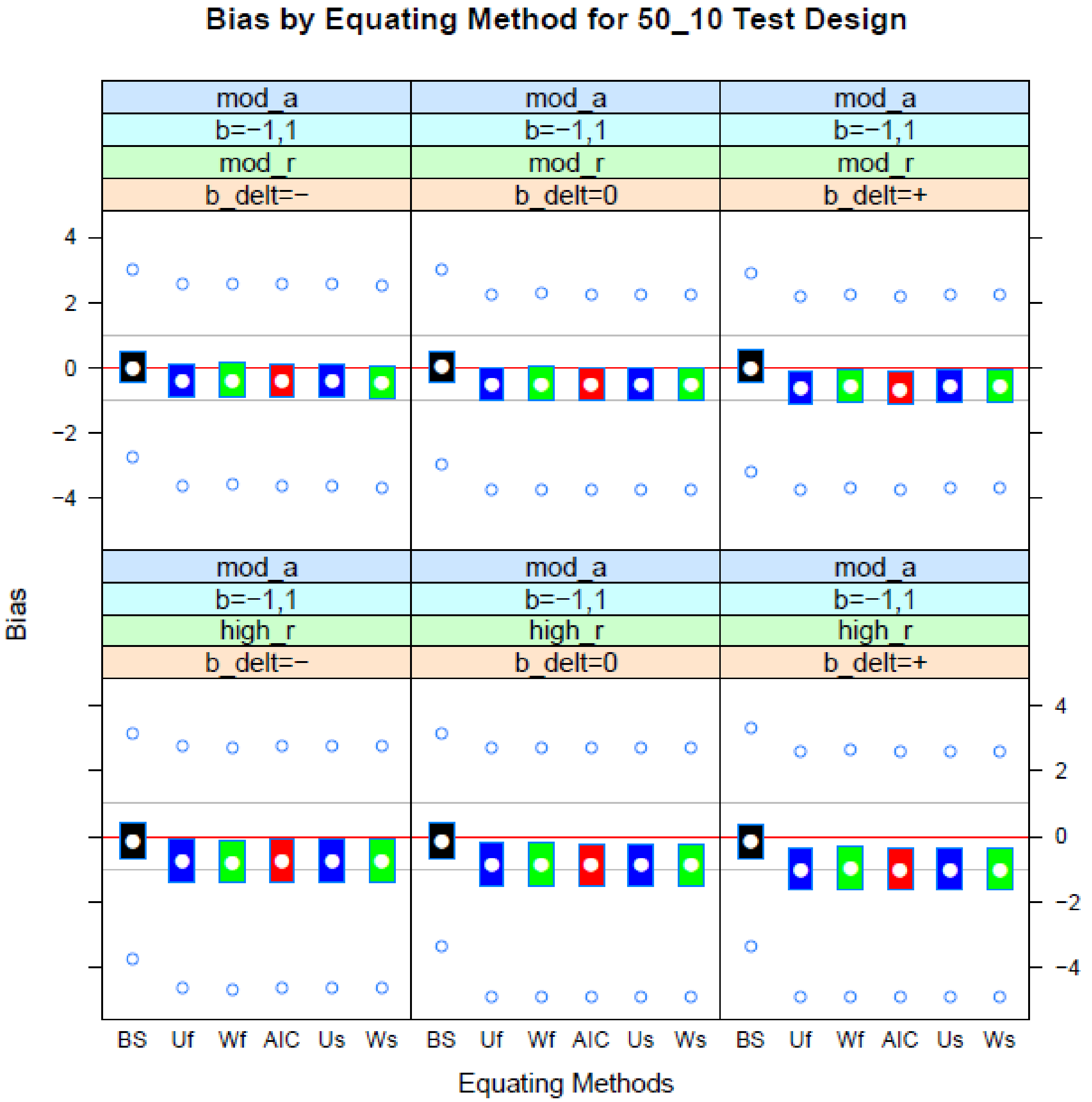


Figure 18. Modified box plot showing bias by equating method. (B = Base form, U = unweighted mean equating, W = weighted mean equating, A = Anchor Item Calibration, U_s = stabilized unweighted mean equating with stabilization and W_s = stabilized weighted mean equating with stabilization)

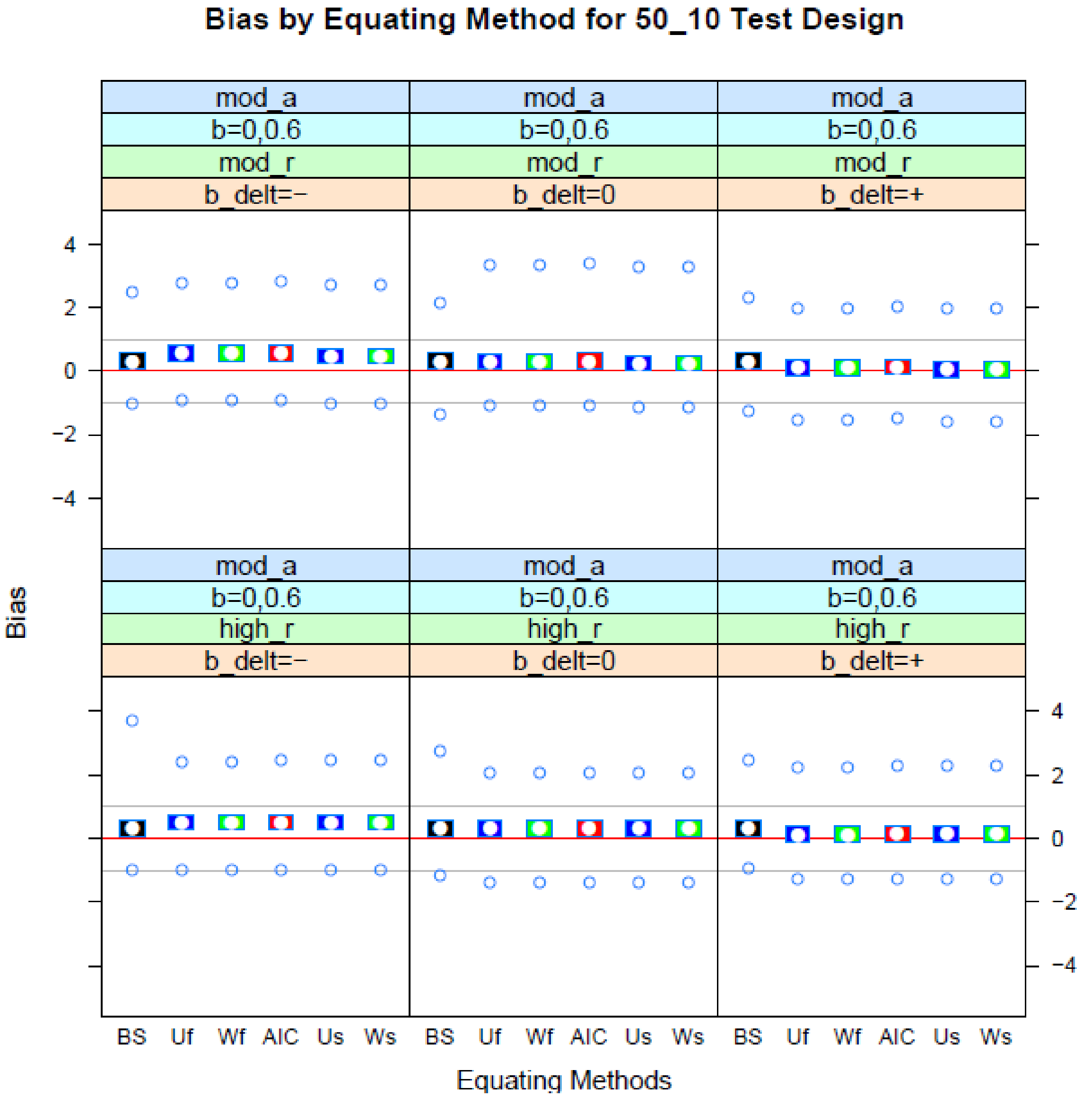


Figure 19. Modified box plot showing bias by equating method. (B = Base form, U = unweighted mean equating, W = weighted mean equating, A = Anchor Item Calibration, U_s = stabilized unweighted mean equating with stabilization and W_s = stabilized weighted mean equating with stabilization)

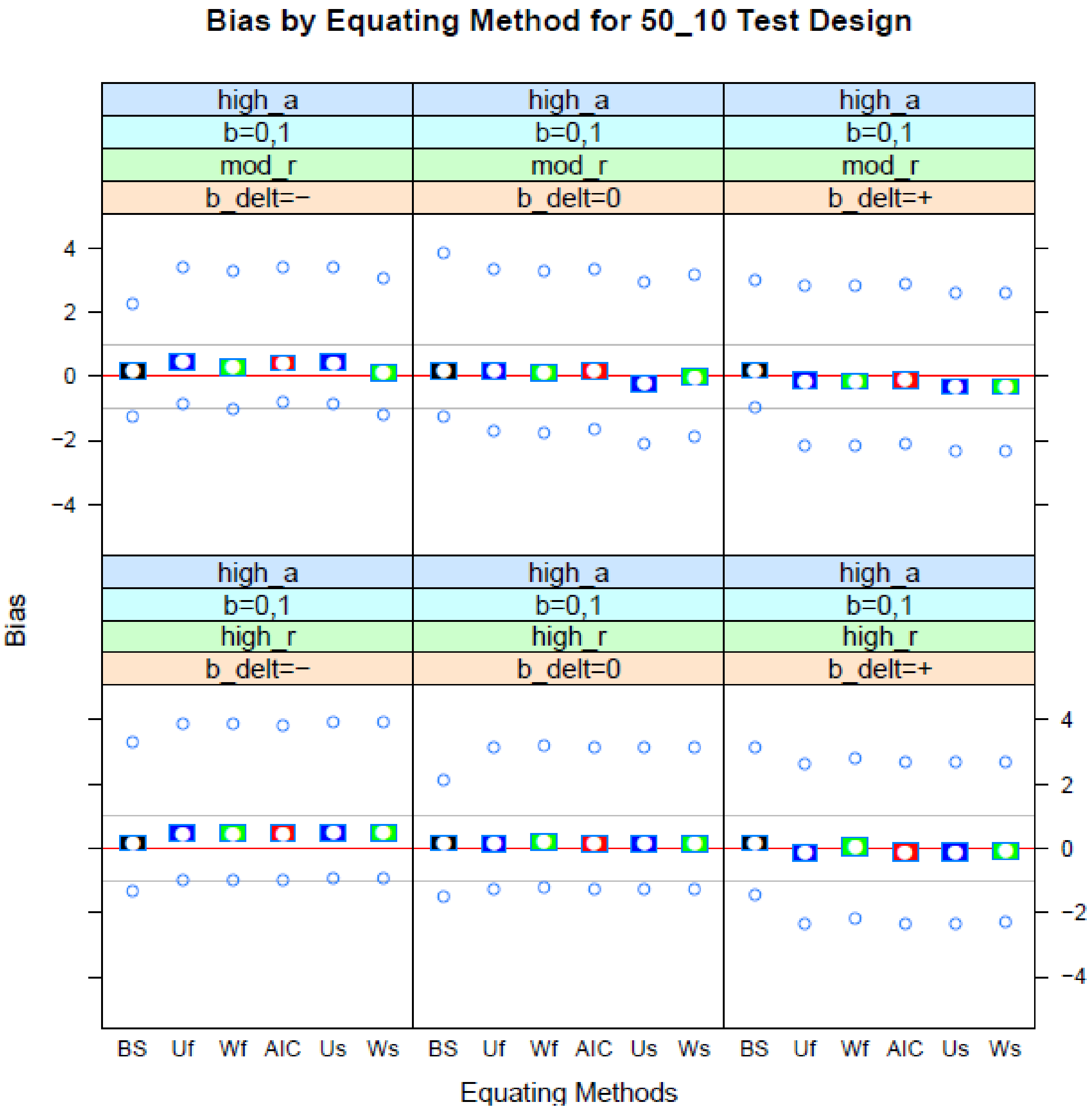


Figure 20. Modified box plot showing bias by equating method. (B = Base form, U = unweighted mean equating, W = weighted mean equating, A = Anchor Item Calibration, U_s = stabilized unweighted mean equating with stabilization and W_s = stabilized weighted mean equating with stabilization)

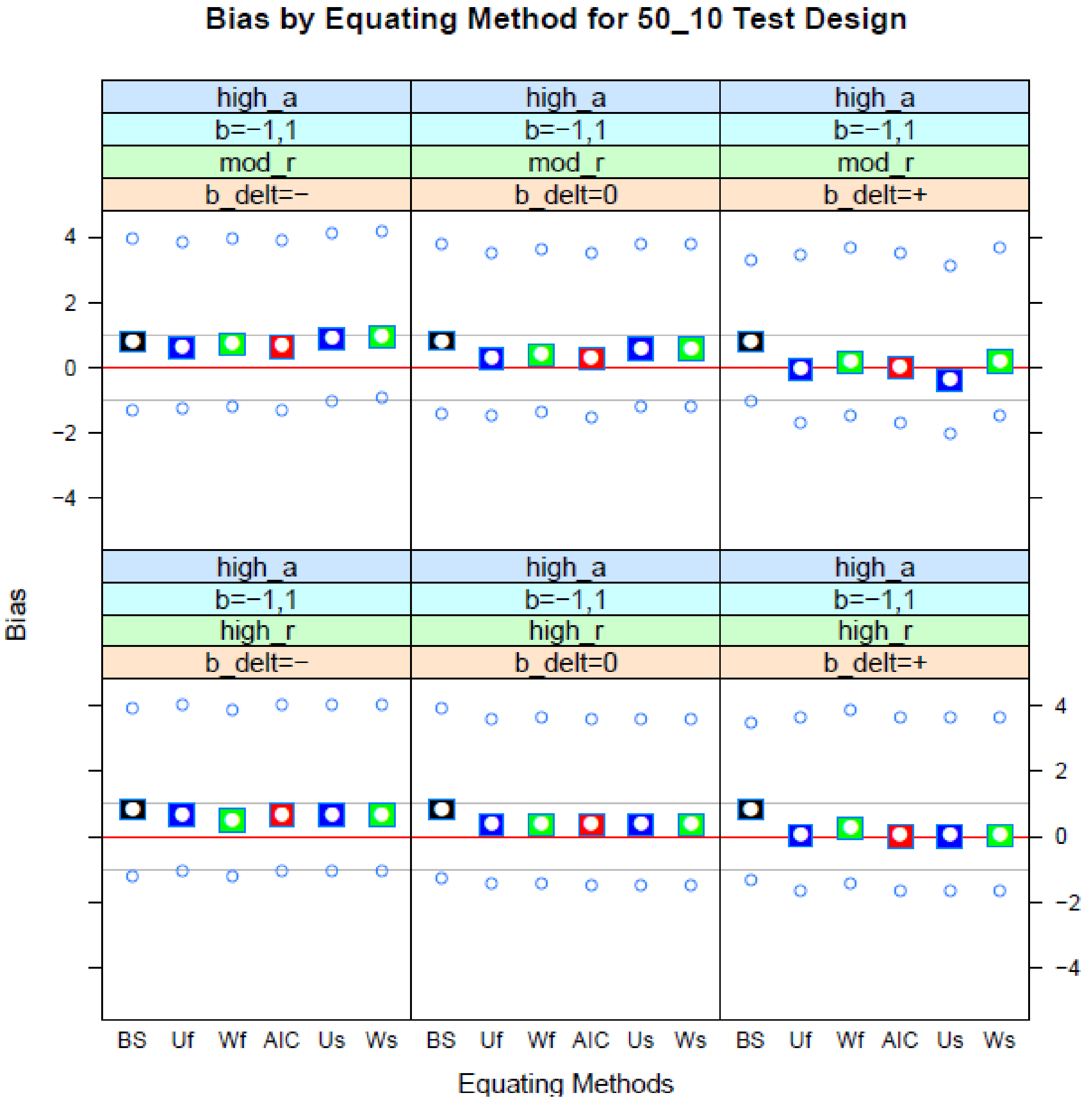


Figure 21. Modified box plot showing bias by equating method. (B = Base form, U = unweighted mean equating, W = weighted mean equating, A = Anchor Item Calibration, U_s = stabilized unweighted mean equating with stabilization and W_s = stabilized weighted mean equating with stabilization)

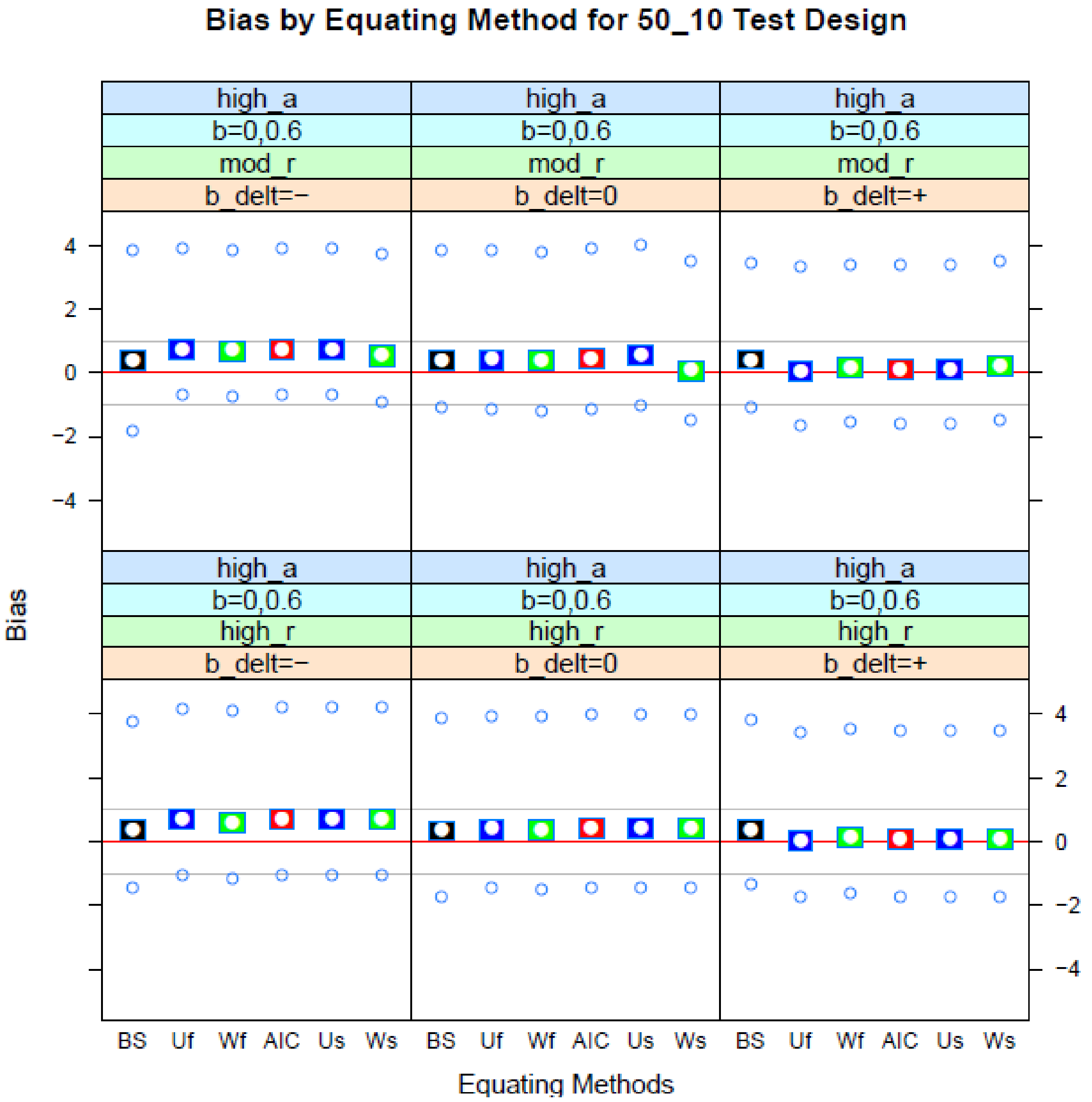


Figure 22. Modified box plot showing bias by equating method. (B = Base form, U = unweighted mean equating, W = weighted mean equating, A = Anchor Item Calibration, U_s = stabilized unweighted mean equating with stabilization and W_s = stabilized weighted mean equating with stabilization)

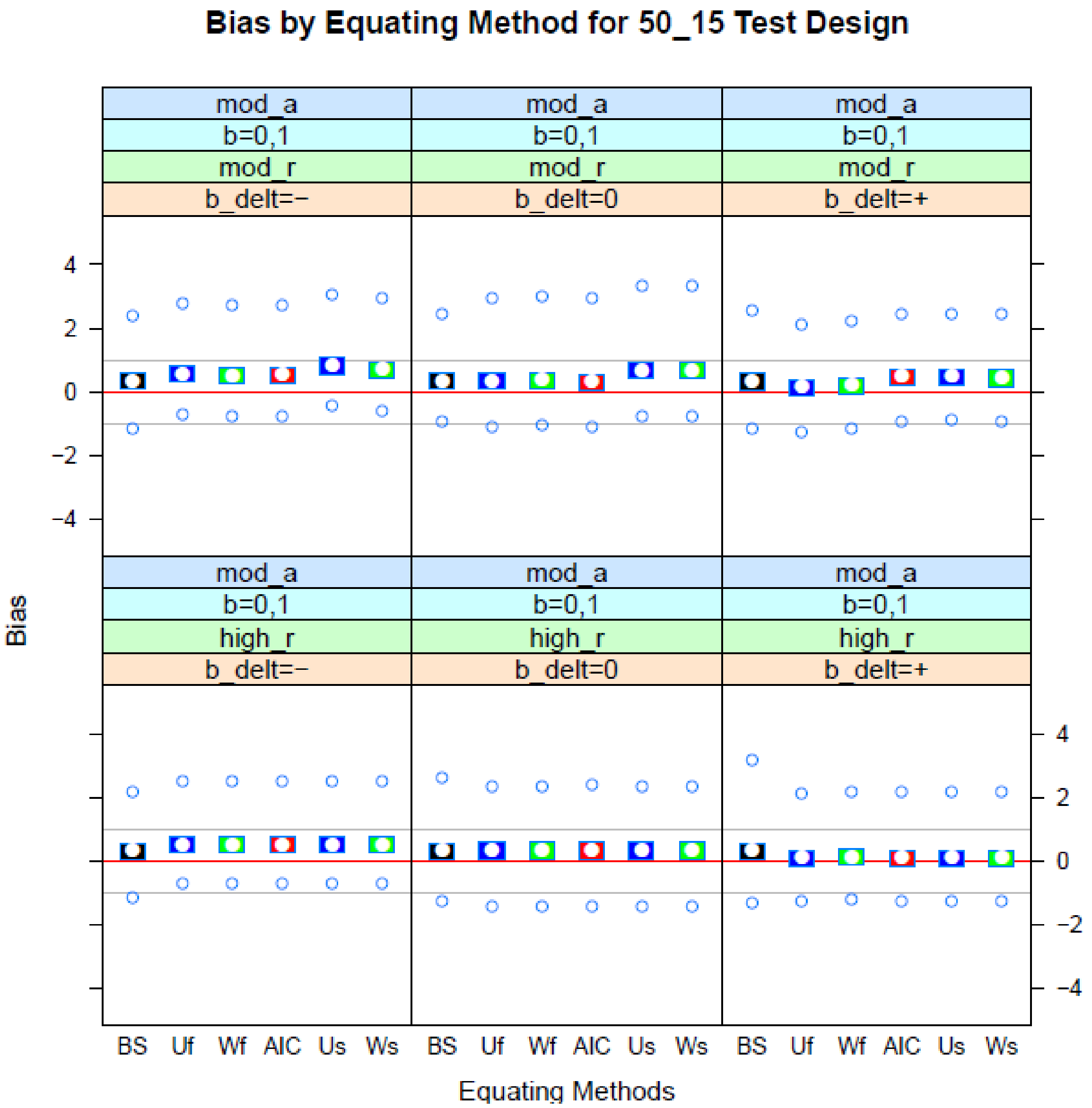


Figure 23. Modified box plot showing bias by equating method. (B = Base form, U = unweighted mean equating, W = weighted mean equating, A = Anchor Item Calibration, U_s = stabilized unweighted mean equating with stabilization and W_s = stabilized weighted mean equating with stabilization)

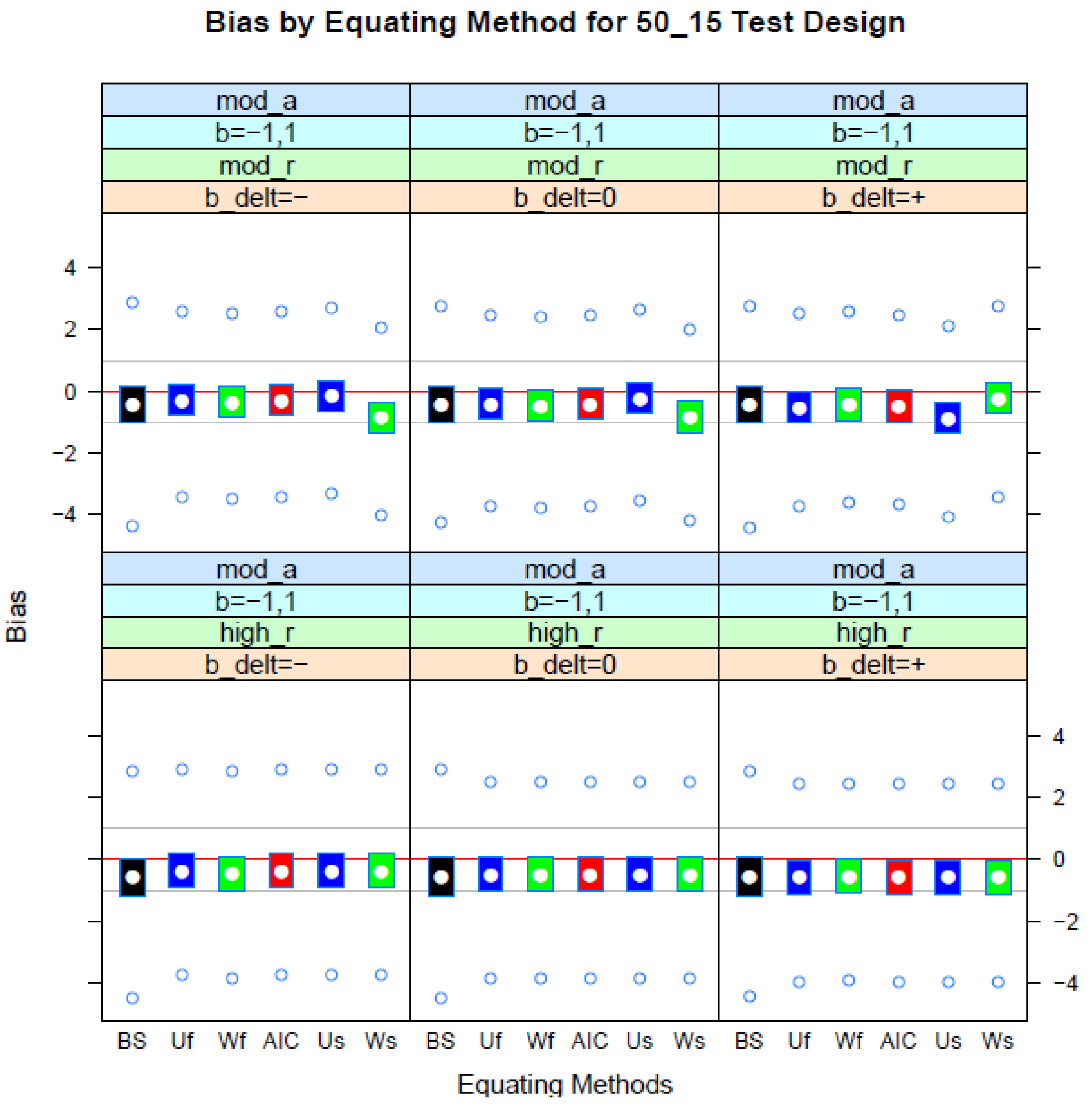


Figure 24. Modified box plot showing bias by equating method. (B = Base form, U = unweighted mean equating, W = weighted mean equating, A = Anchor Item Calibration, U_s = stabilized unweighted mean equating with stabilization and W_s = stabilized weighted mean equating with stabilization

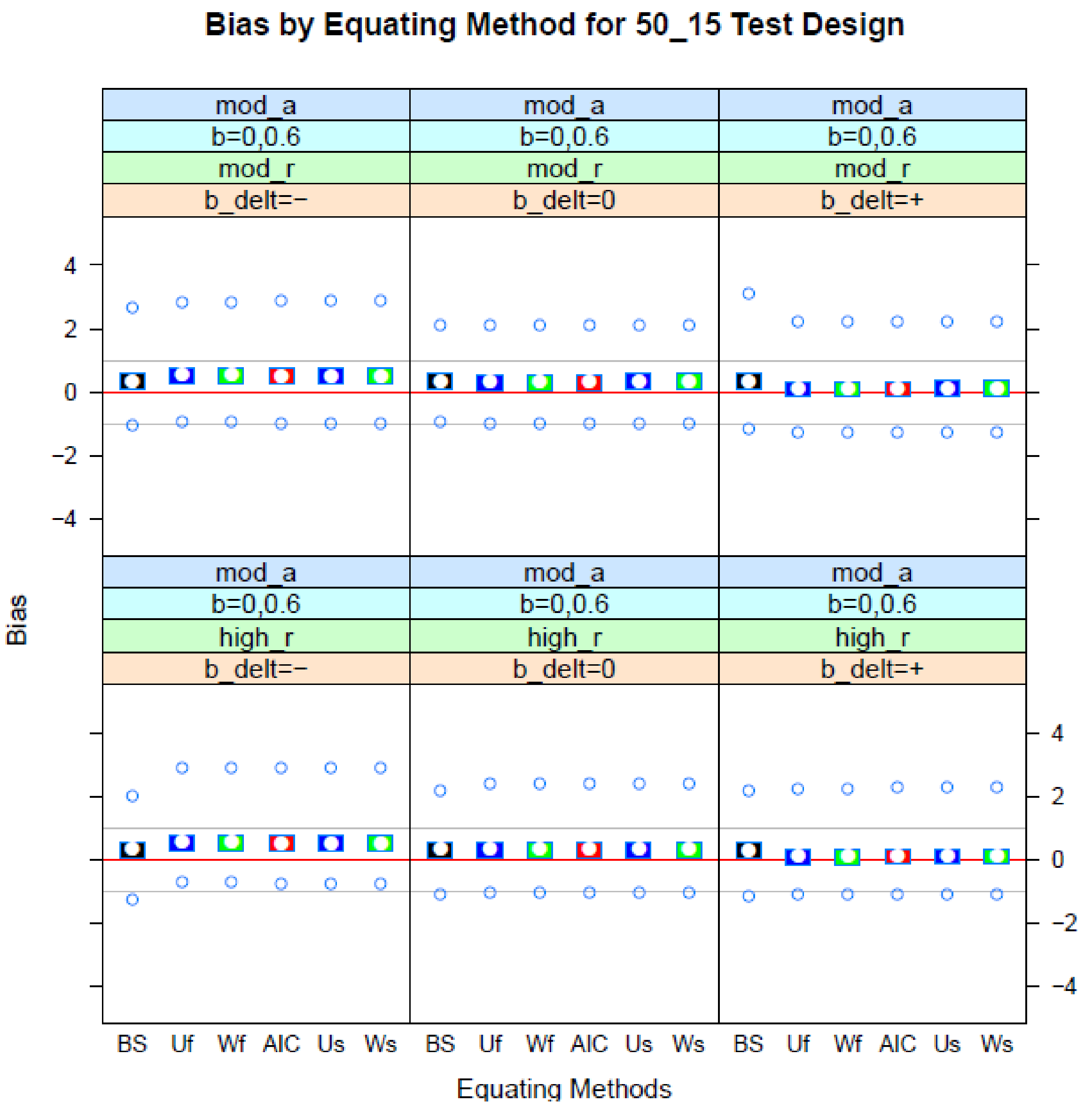


Figure 25. Modified box plot showing bias by equating method. (B = Base form, U = unweighted mean equating, W = weighted mean equating, A = Anchor Item Calibration, U_s = stabilized unweighted mean equating with stabilization and W_s = stabilized weighted mean equating with stabilization)

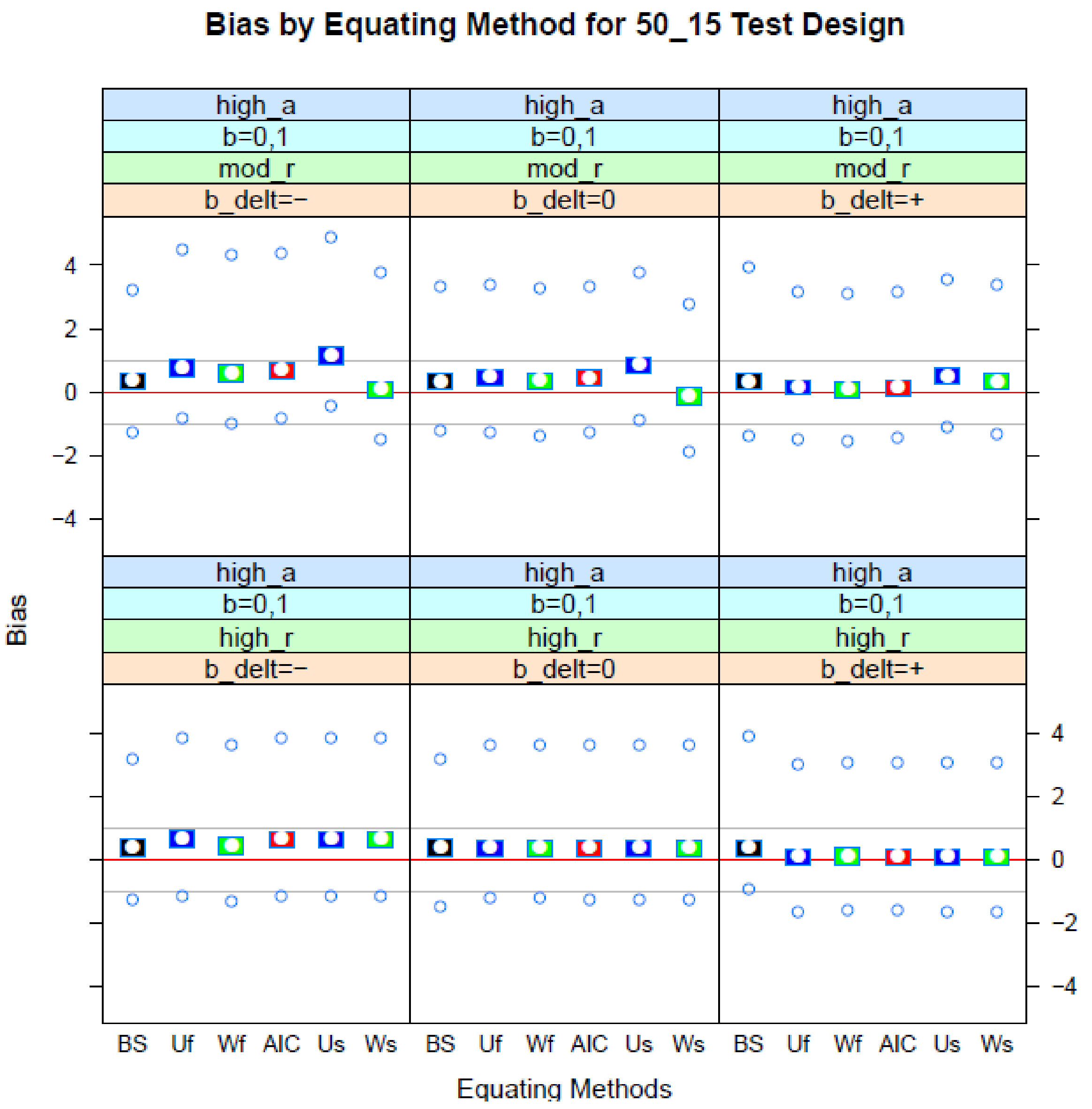


Figure 26. Modified box plot showing bias by equating method. (B = Base form, U = unweighted mean equating, W = weighted mean equating, A = Anchor Item Calibration, U_s = stabilized unweighted mean equating with stabilization and W_s = stabilized weighted mean equating with stabilization)

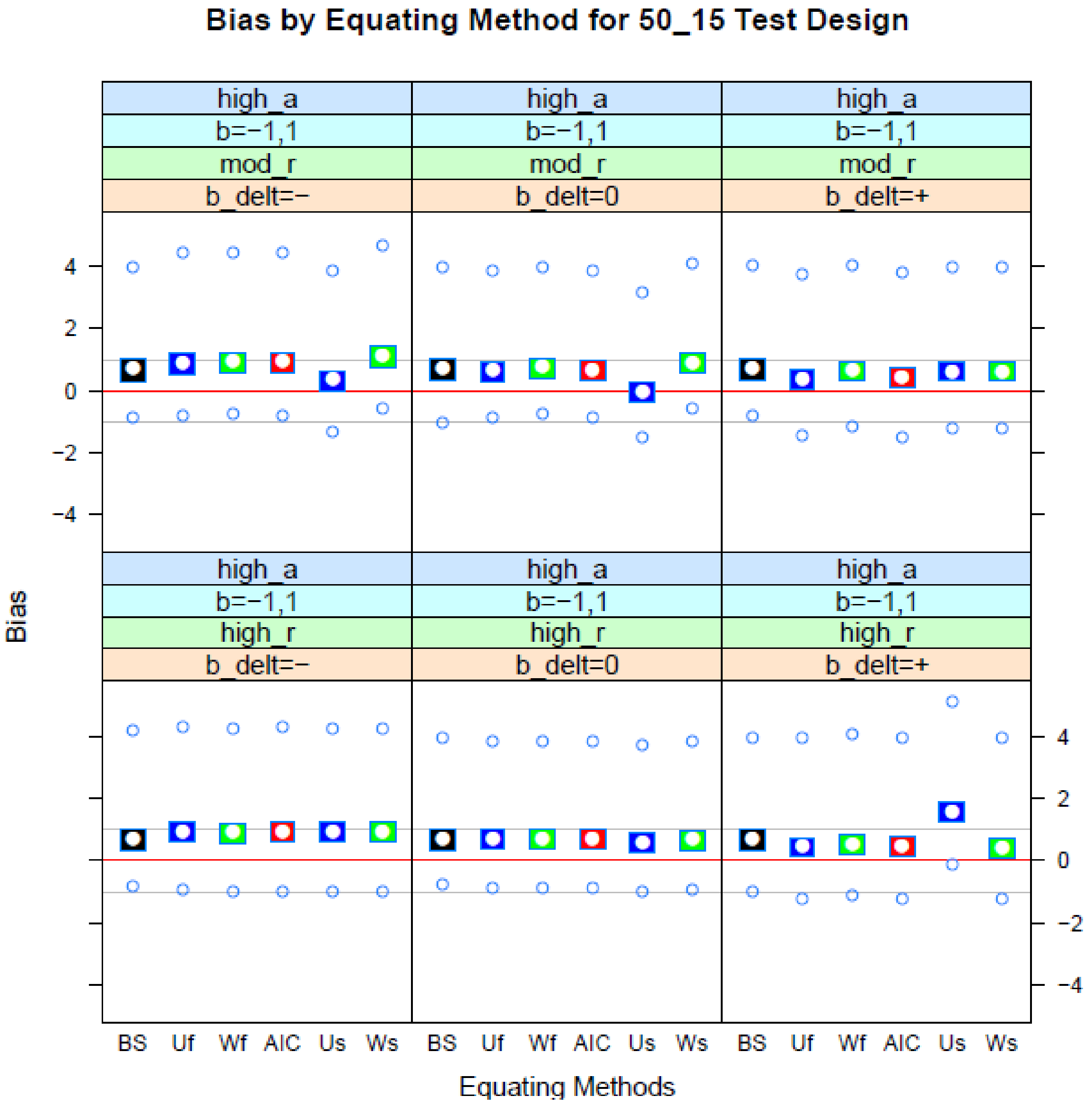


Figure 27. Modified box plot showing bias by equating method. (B = Base form, U = unweighted mean equating, W = weighted mean equating, A = Anchor Item Calibration, U_s = stabilized unweighted mean equating with stabilization and W_s = stabilized weighted mean equating with stabilization)

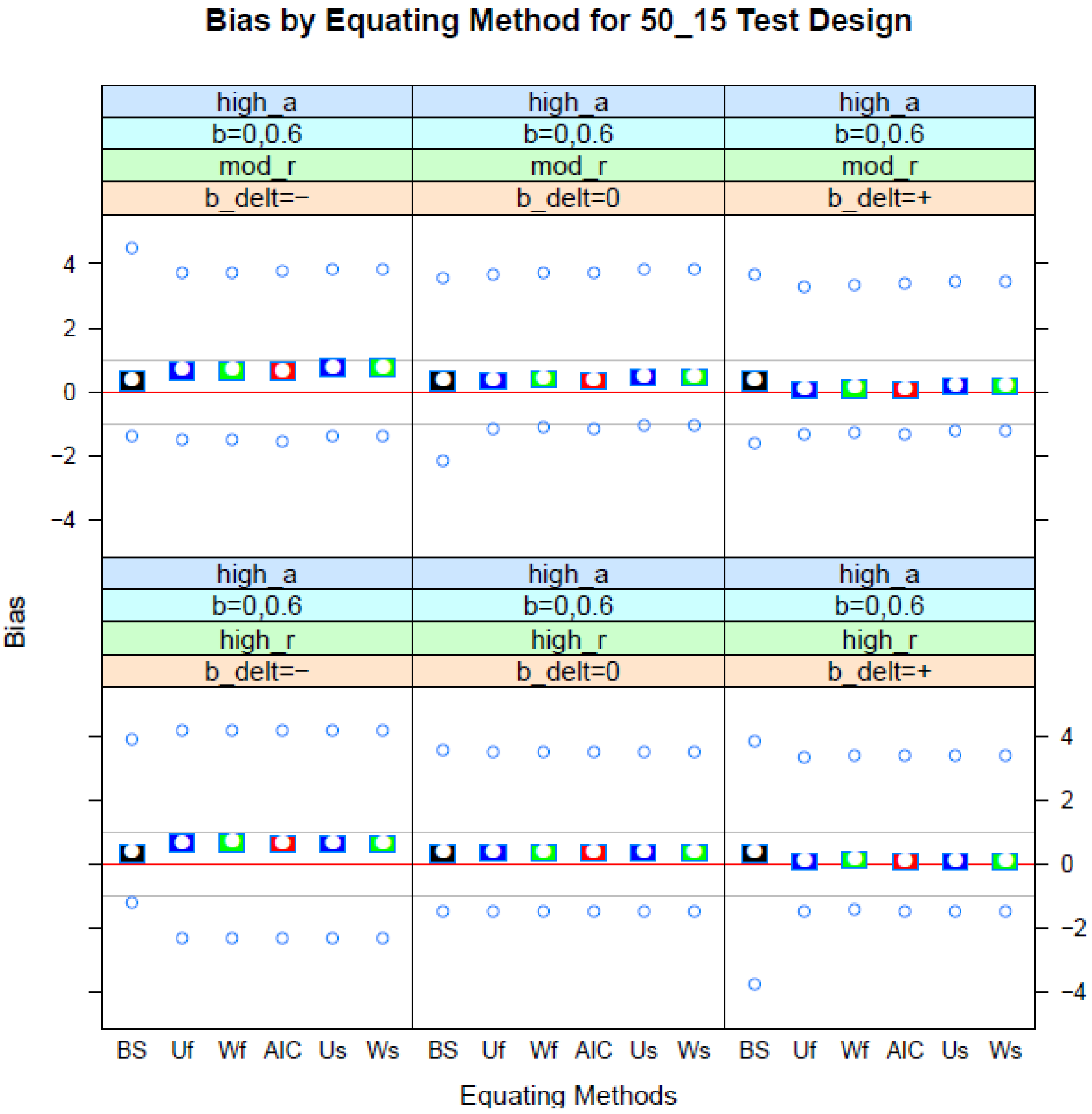


Figure 28. Modified box plot showing bias by equating method. (B = Base form, U = unweighted mean equating, W = weighted mean equating, A = Anchor Item Calibration, U_s = stabilized unweighted mean equating with stabilization and W_s = stabilized weighted mean equating with stabilization)

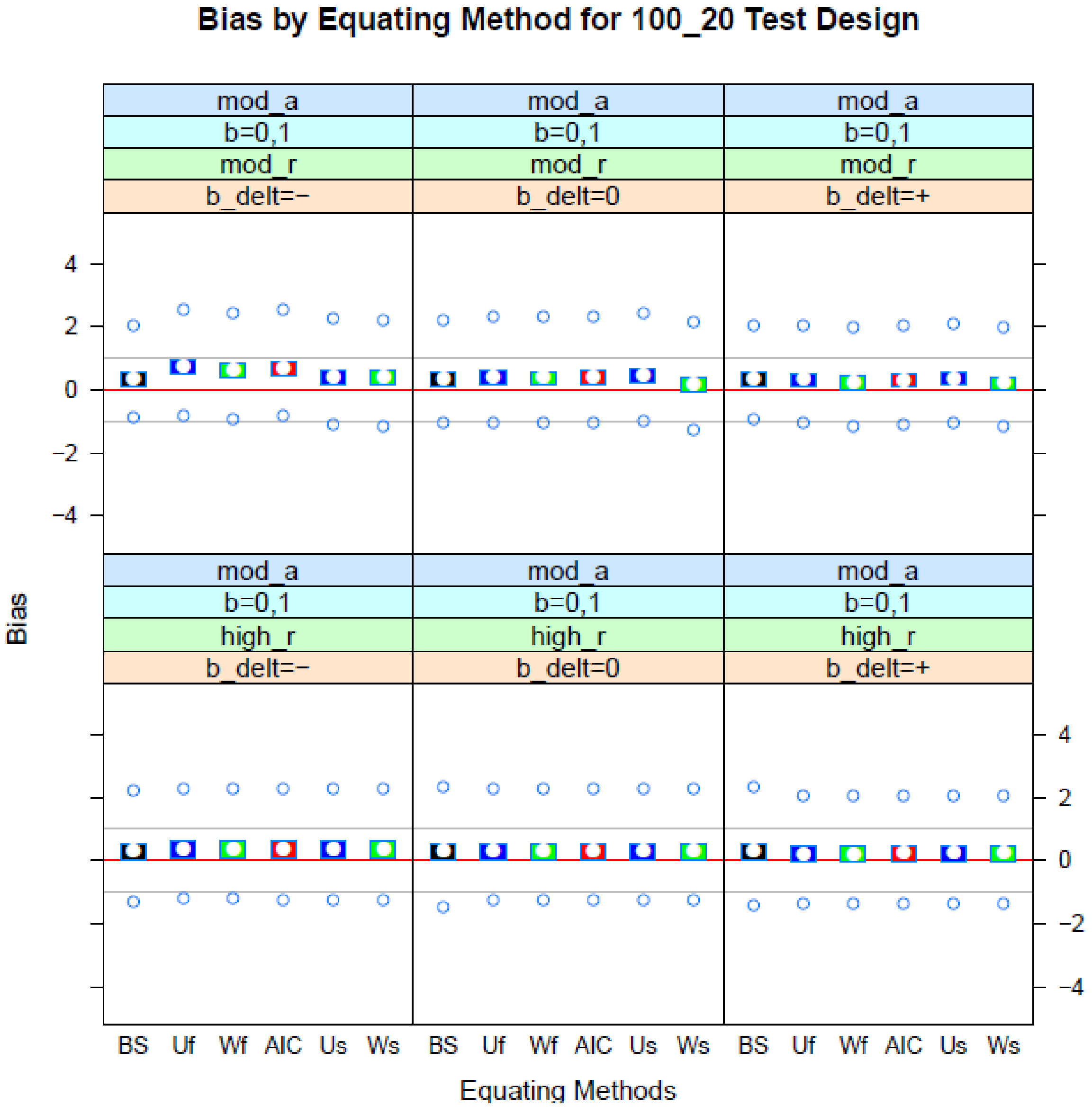


Figure 29. Modified box plot showing bias by equating method. (B = Base form, U = unweighted mean equating, W = weighted mean equating, A = Anchor Item Calibration, U_s = stabilized unweighted mean equating with stabilization and W_s = stabilized weighted mean equating with stabilization)

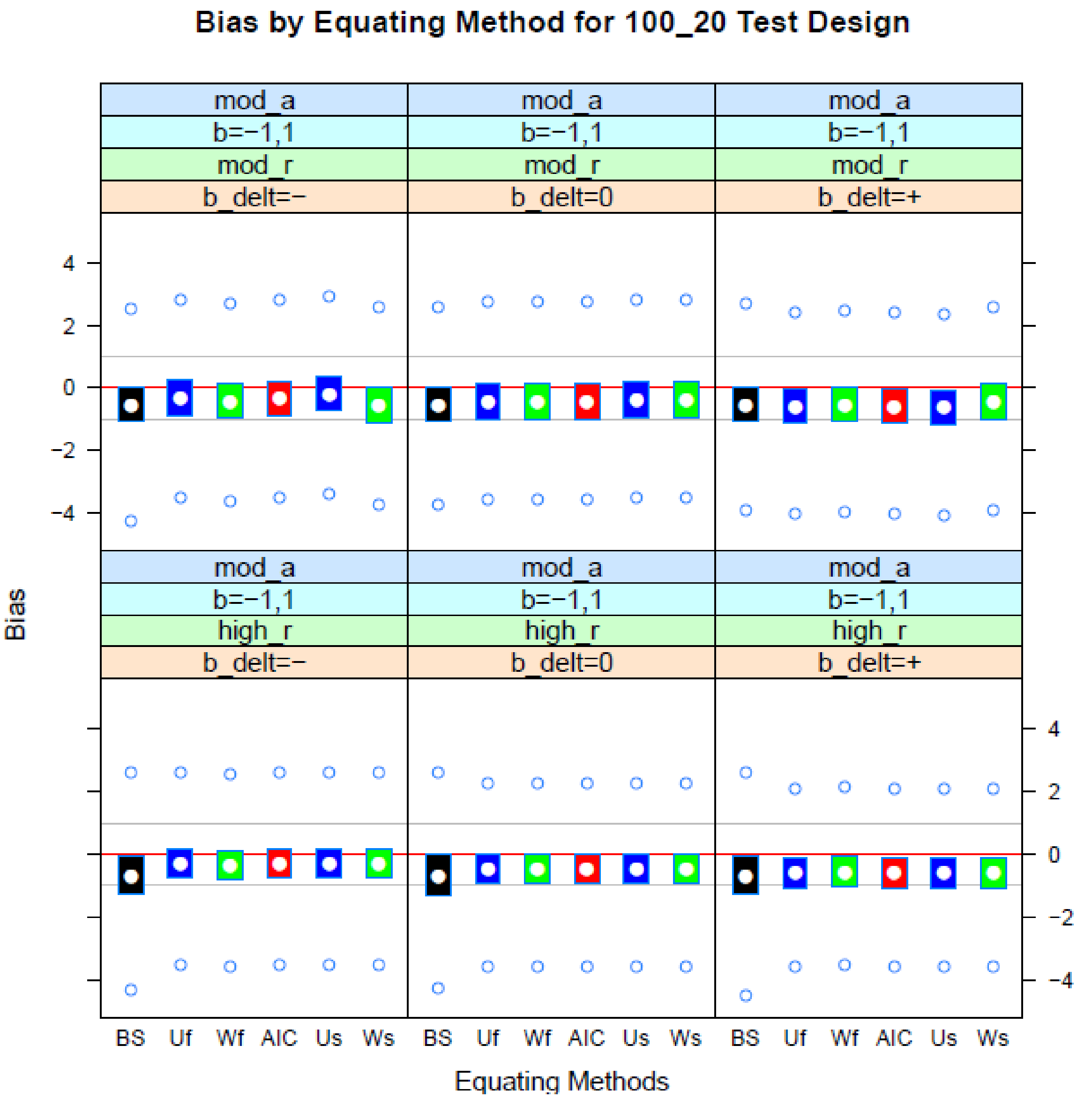


Figure 30. Modified box plot showing bias by equating method. (B = Base form, U = unweighted mean equating, W = weighted mean equating, A = Anchor Item Calibration, U_s = stabilized unweighted mean equating with stabilization and W_s = stabilized weighted mean equating with stabilization)

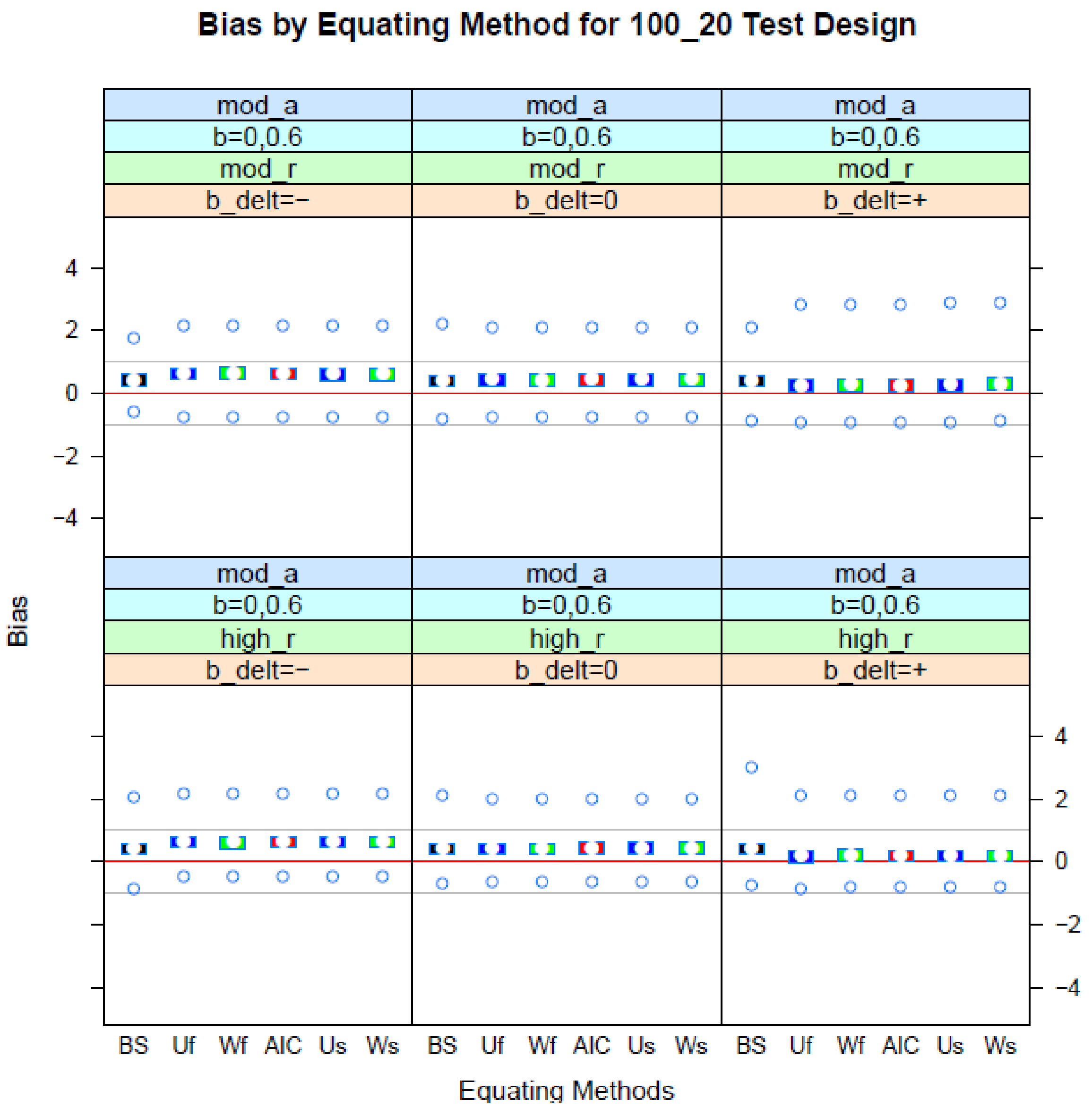


Figure 31. Modified box plot showing bias by equating method. (B = Base form, U = unweighted mean equating, W = weighted mean equating, A = Anchor Item Calibration, U_s = stabilized unweighted mean equating with stabilization and W_s = stabilized weighted mean equating with stabilization)

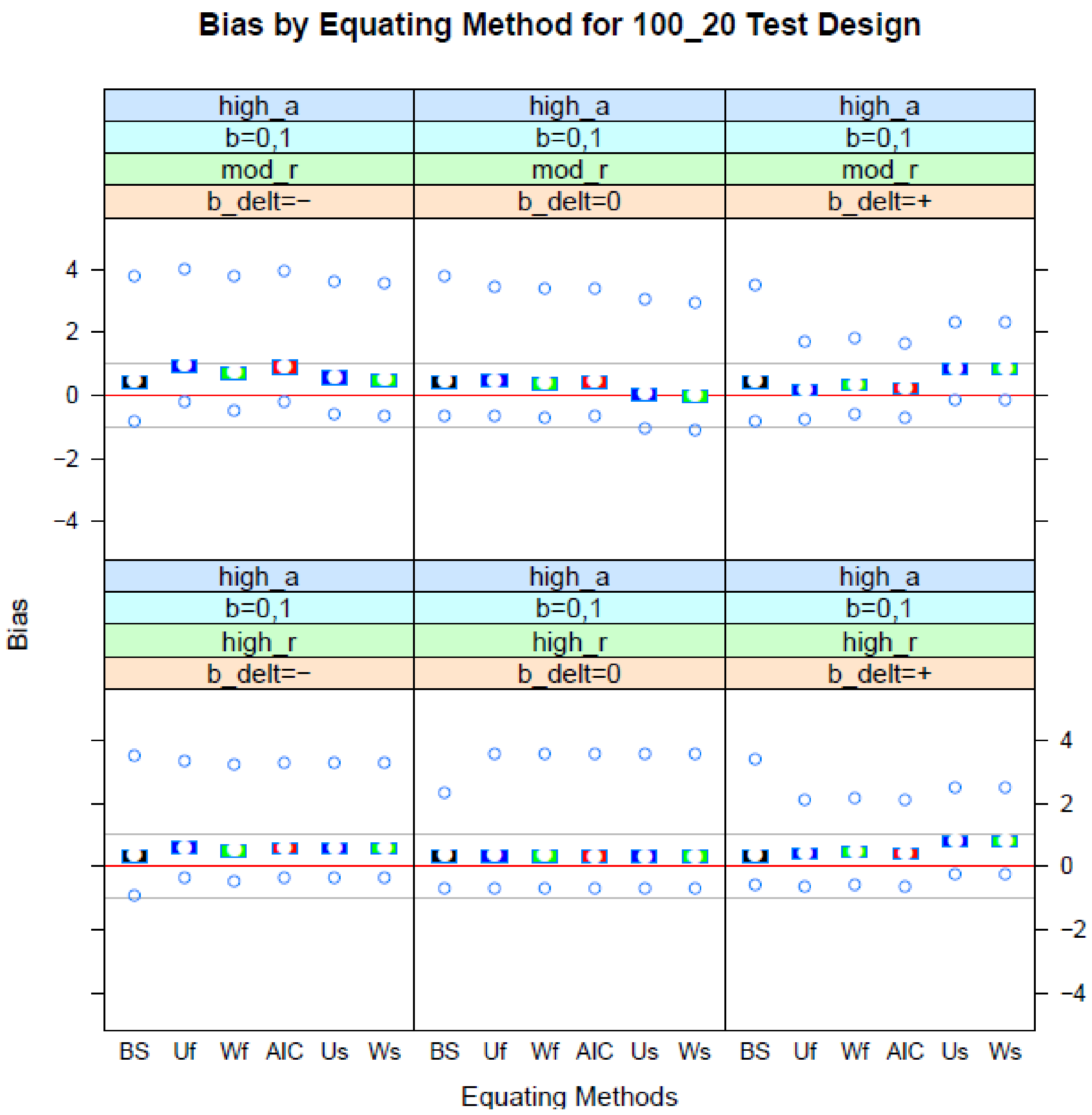


Figure 32. Modified box plot showing bias by equating method. (B = Base form, U = unweighted mean equating, W = weighted mean equating, A = Anchor Item Calibration, U_s = stabilized unweighted mean equating with stabilization and W_s = stabilized weighted mean equating with stabilization)

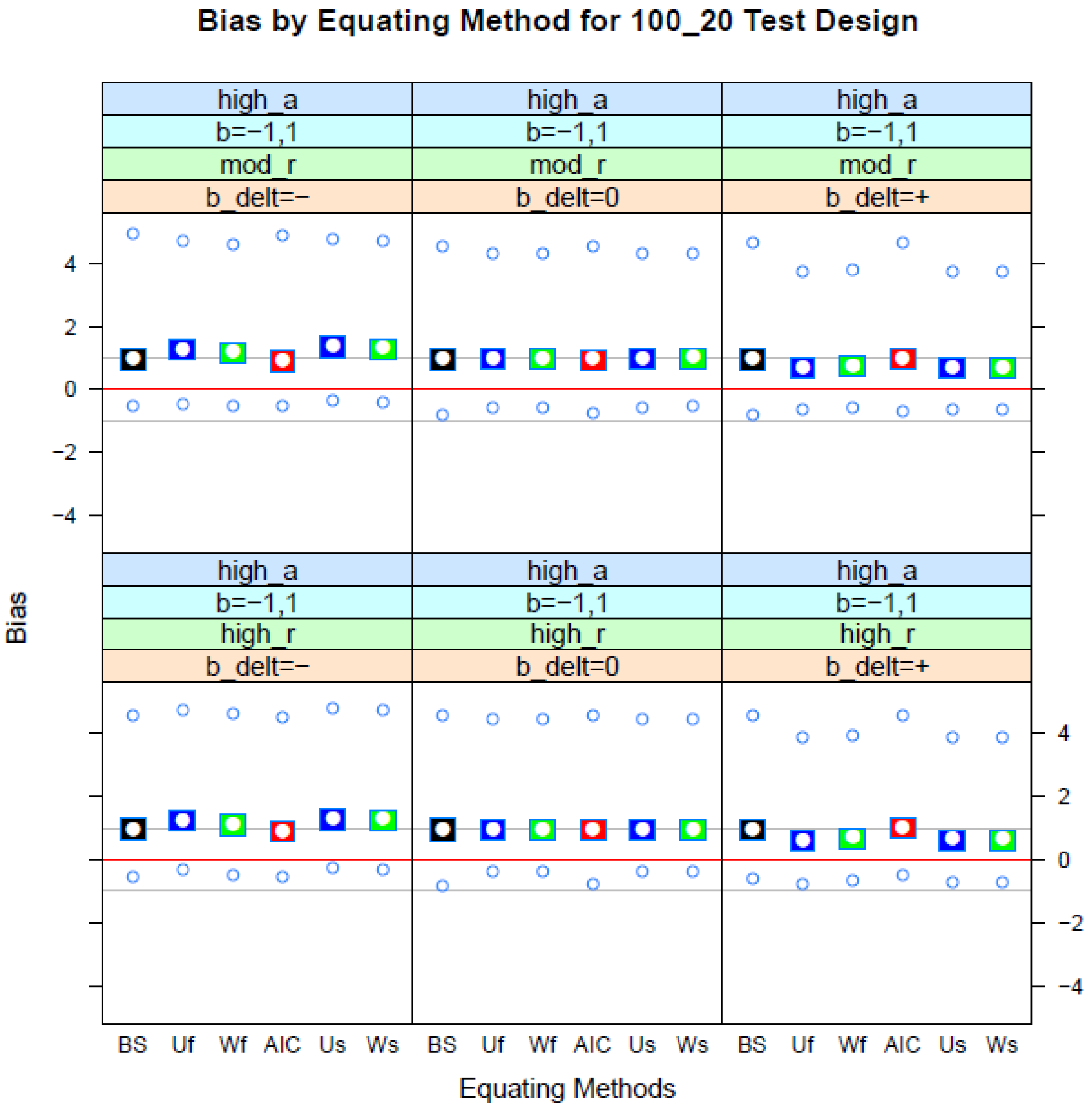


Figure 33. Modified box plot showing bias by equating method. (B = Base form, U = unweighted mean equating, W = weighted mean equating, A = Anchor Item Calibration, U_s = stabilized unweighted mean equating with stabilization and W_s = stabilized weighted mean equating with stabilization)

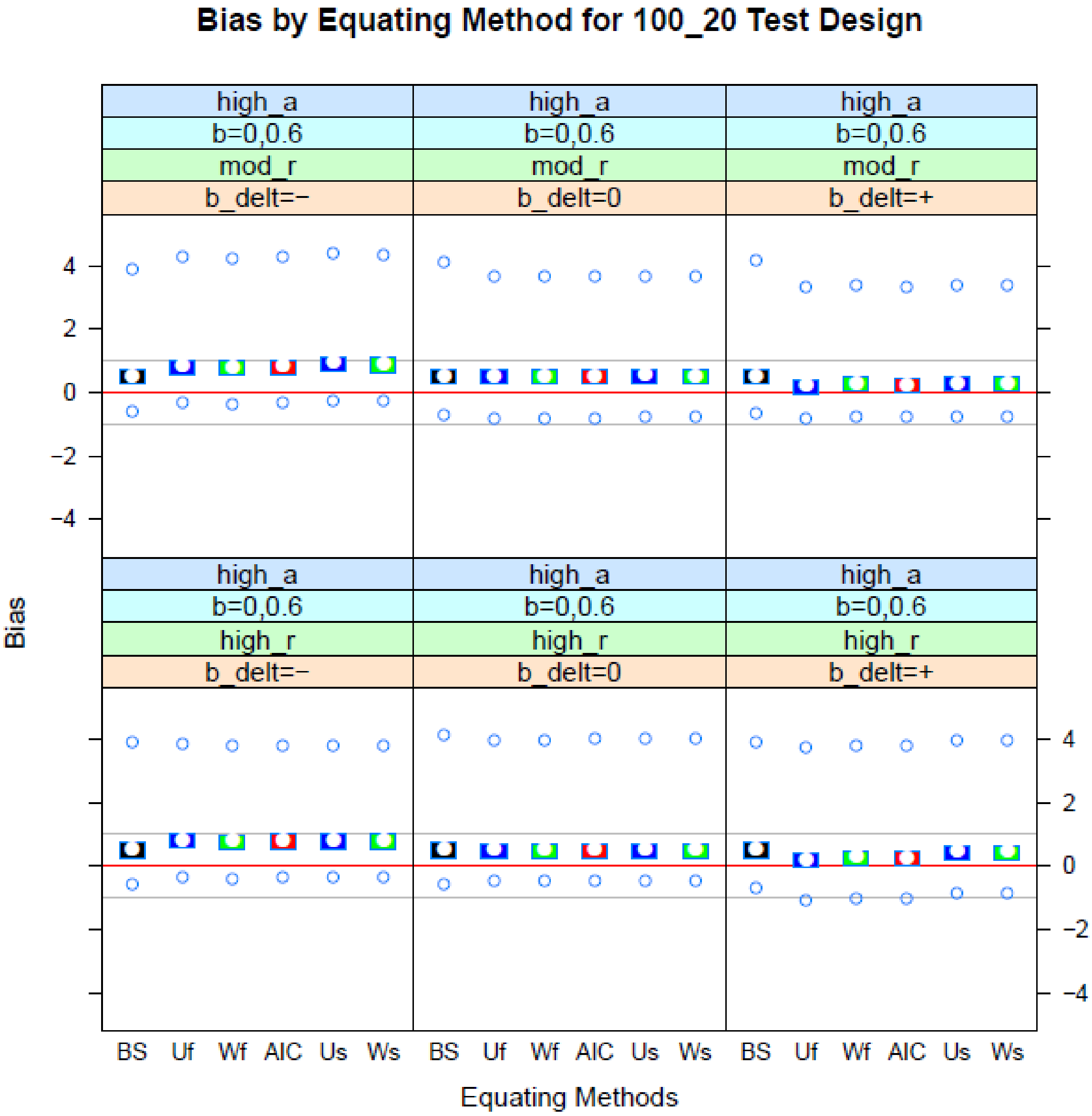


Figure 34. Modified box plot showing bias by equating method. (B = Base form, U = unweighted mean equating, W = weighted mean equating, A = Anchor Item Calibration, U_s = stabilized unweighted mean equating with stabilization and W_s = stabilized weighted mean equating with stabilization)

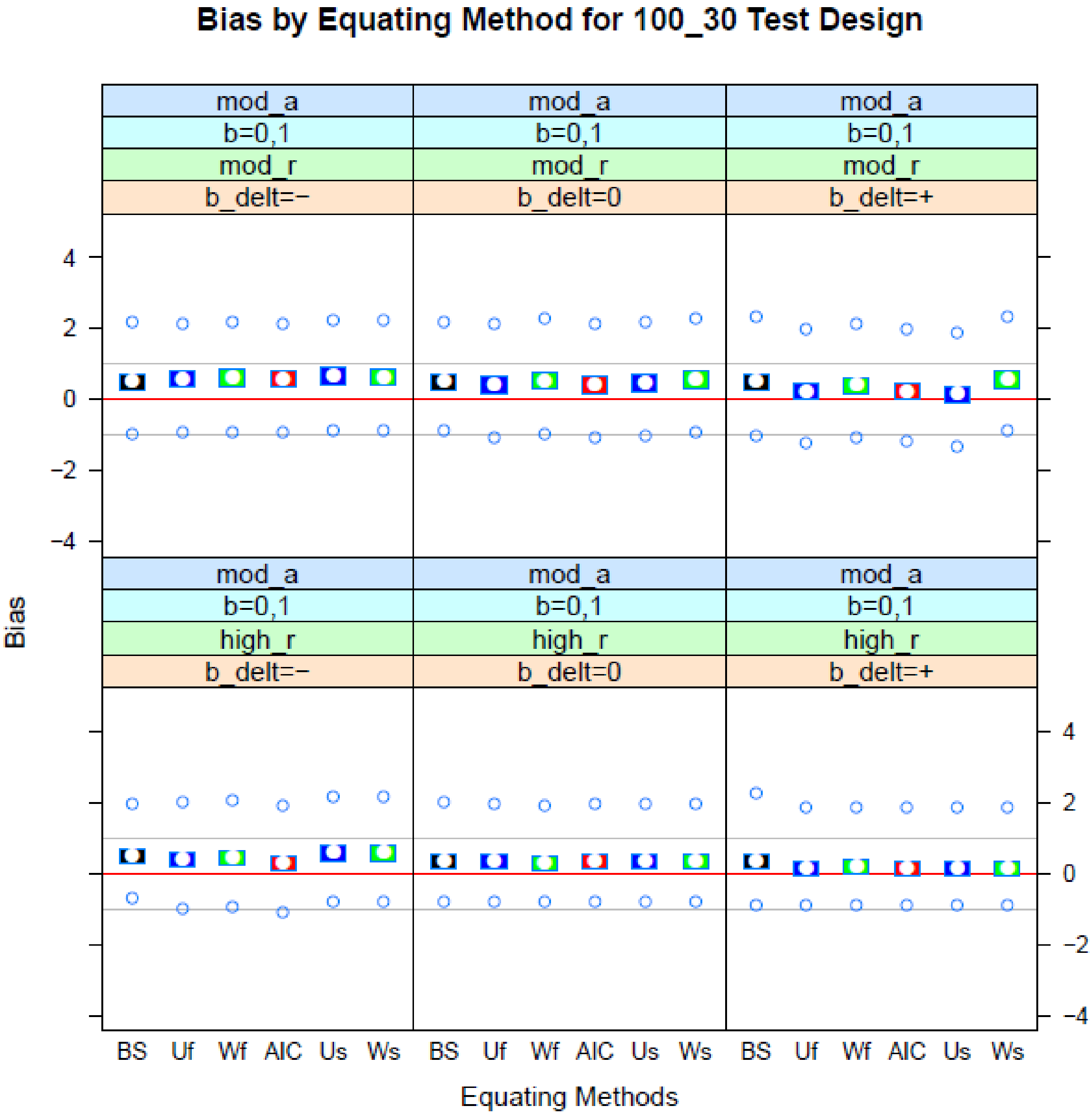


Figure 35. Modified box plot showing bias by equating method. (B = Base form, U = unweighted mean equating, W = weighted mean equating, A = Anchor Item Calibration, U_s = stabilized unweighted mean equating with stabilization and W_s = stabilized weighted mean equating with stabilization)

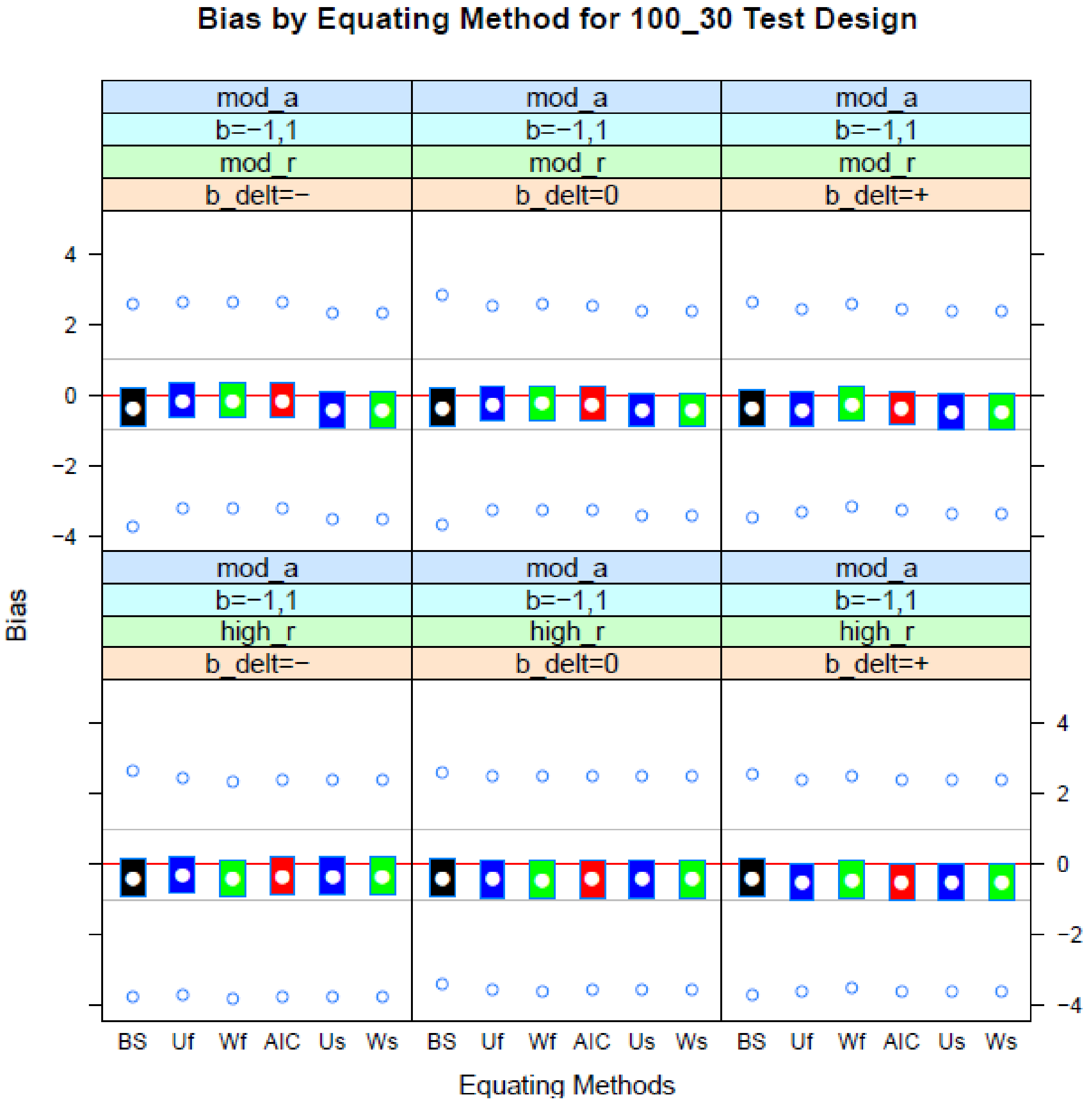


Figure 36. Modified box plot showing bias by equating method. (B = Base form, U = unweighted mean equating, W = weighted mean equating, A = Anchor Item Calibration, U_s = stabilized unweighted mean equating with stabilization and W_s = stabilized weighted mean equating with stabilization)

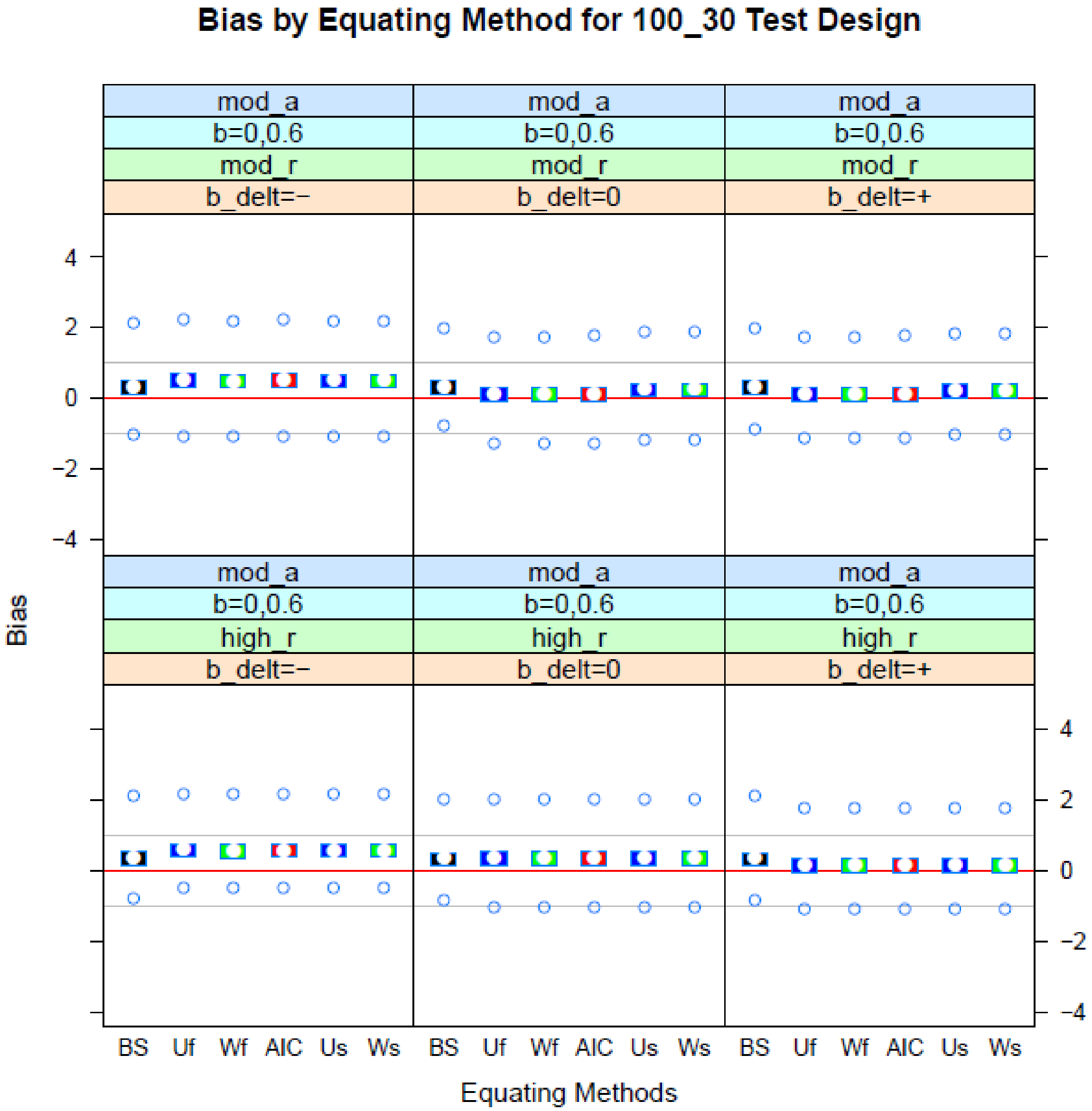


Figure 37. Modified box plot showing bias by equating method. (B = Base form, U = unweighted mean equating, W = weighted mean equating, A = Anchor Item Calibration, U_s = stabilized unweighted mean equating with stabilization and W_s = stabilized weighted mean equating with stabilization)

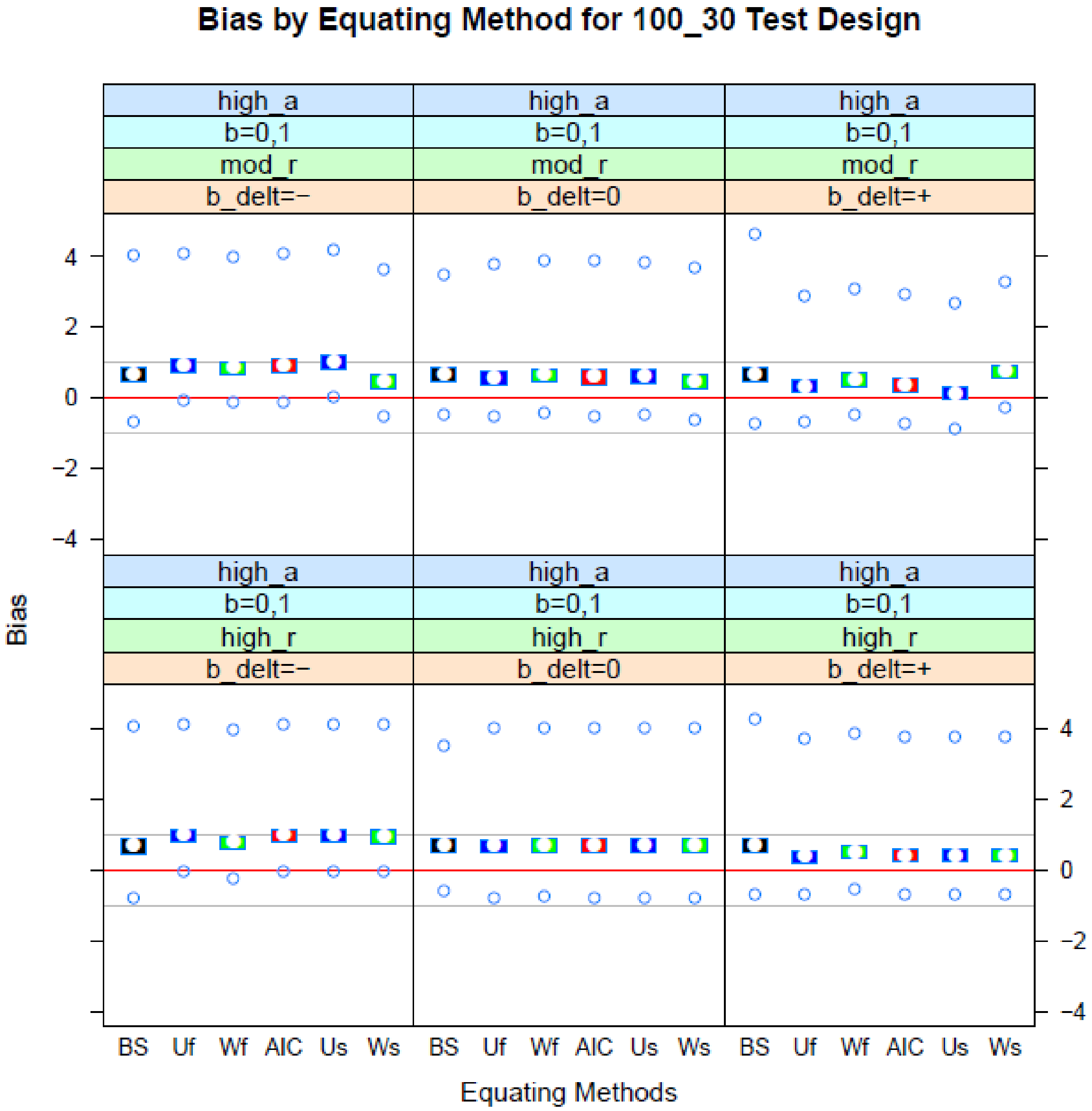


Figure 38. Modified box plot showing bias by equating method. (B = Base form, U = unweighted mean equating, W = weighted mean equating, A = Anchor Item Calibration, U_s = stabilized unweighted mean equating with stabilization and W_s = stabilized weighted mean equating with stabilization)

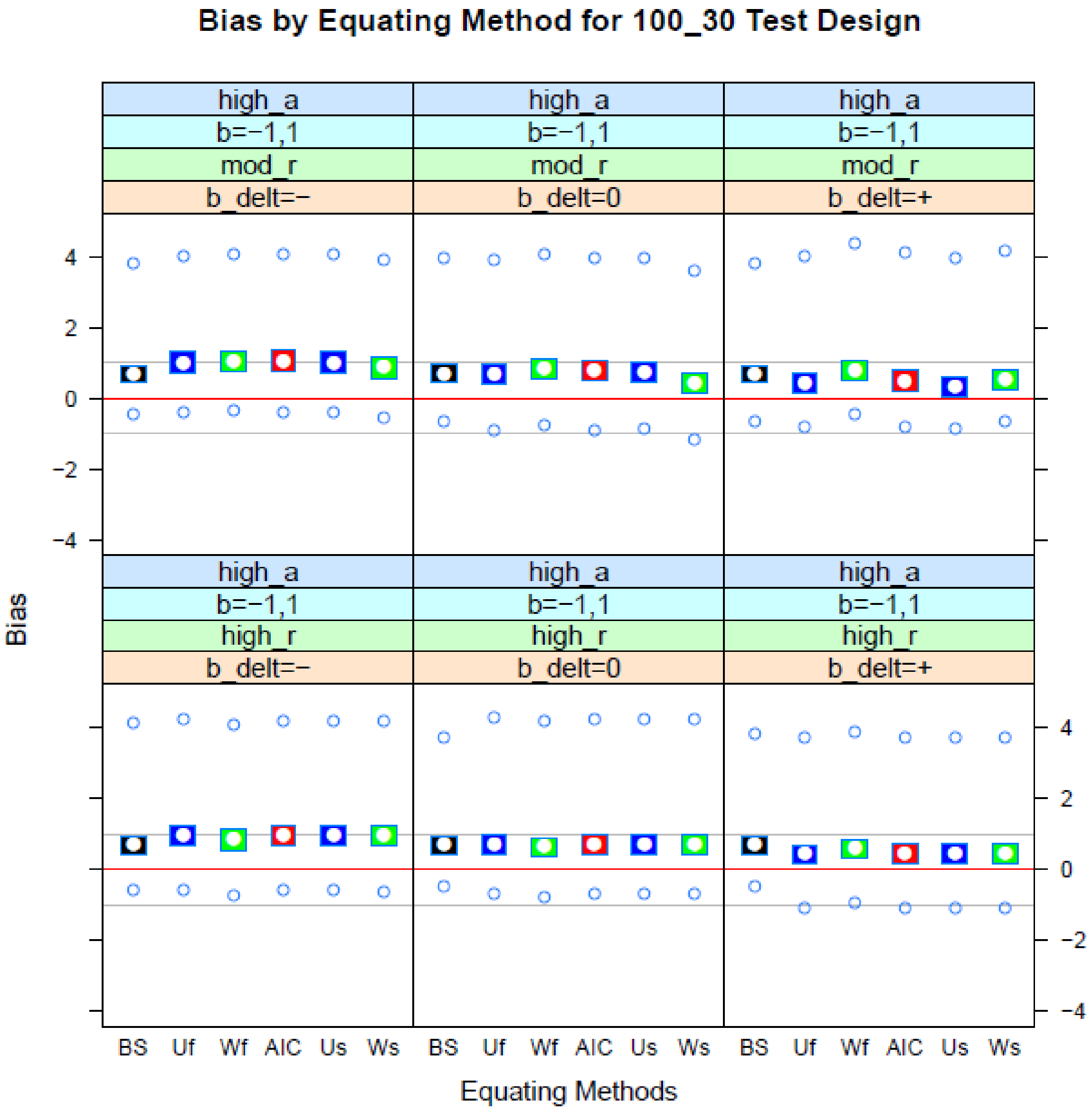


Figure 39. Modified box plot showing bias by equating method. (B = Base form, U = unweighted mean equating, W = weighted mean equating, A = Anchor Item Calibration, U_s = stabilized unweighted mean equating with stabilization and W_s = stabilized weighted mean equating with stabilization)

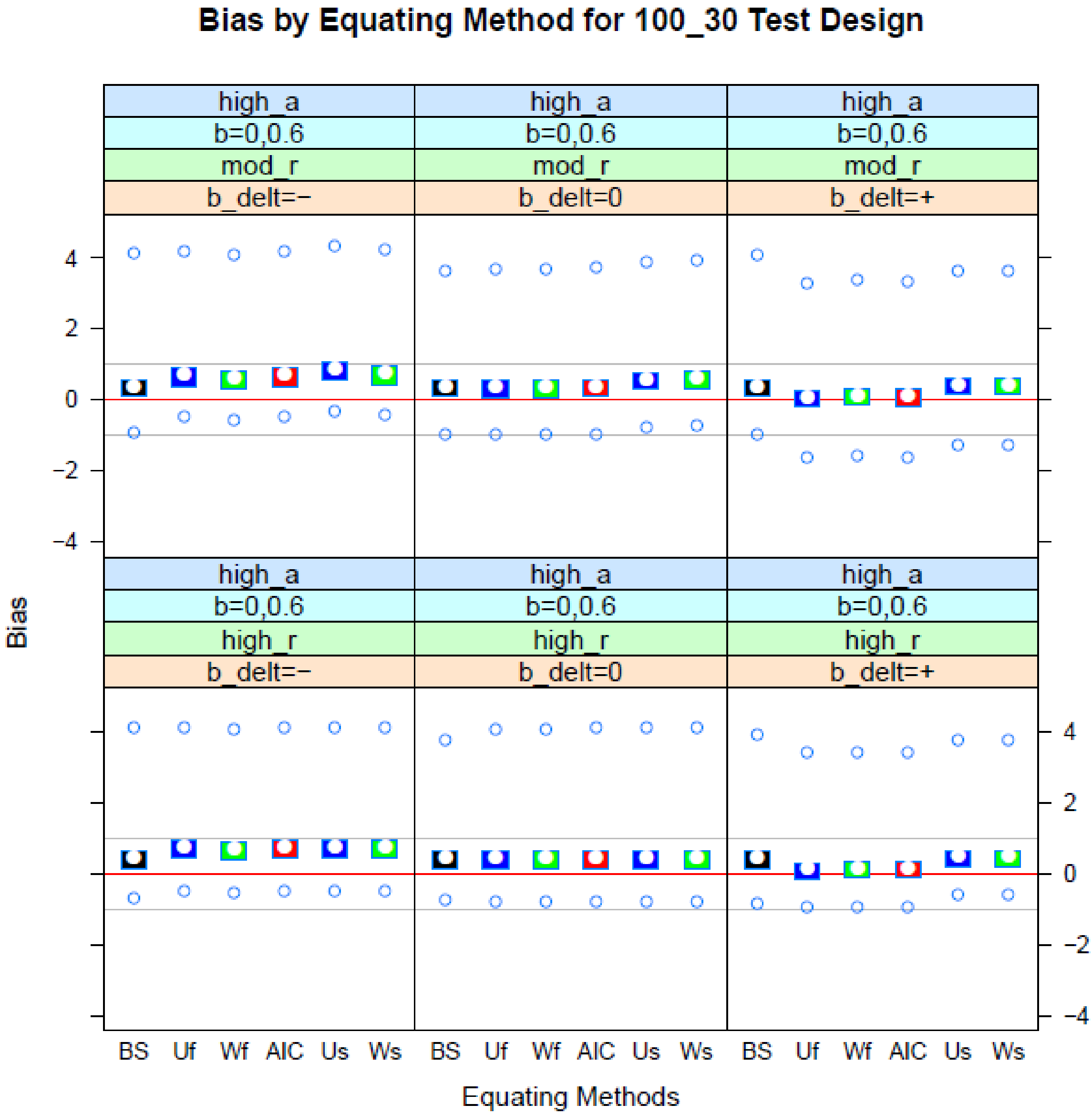
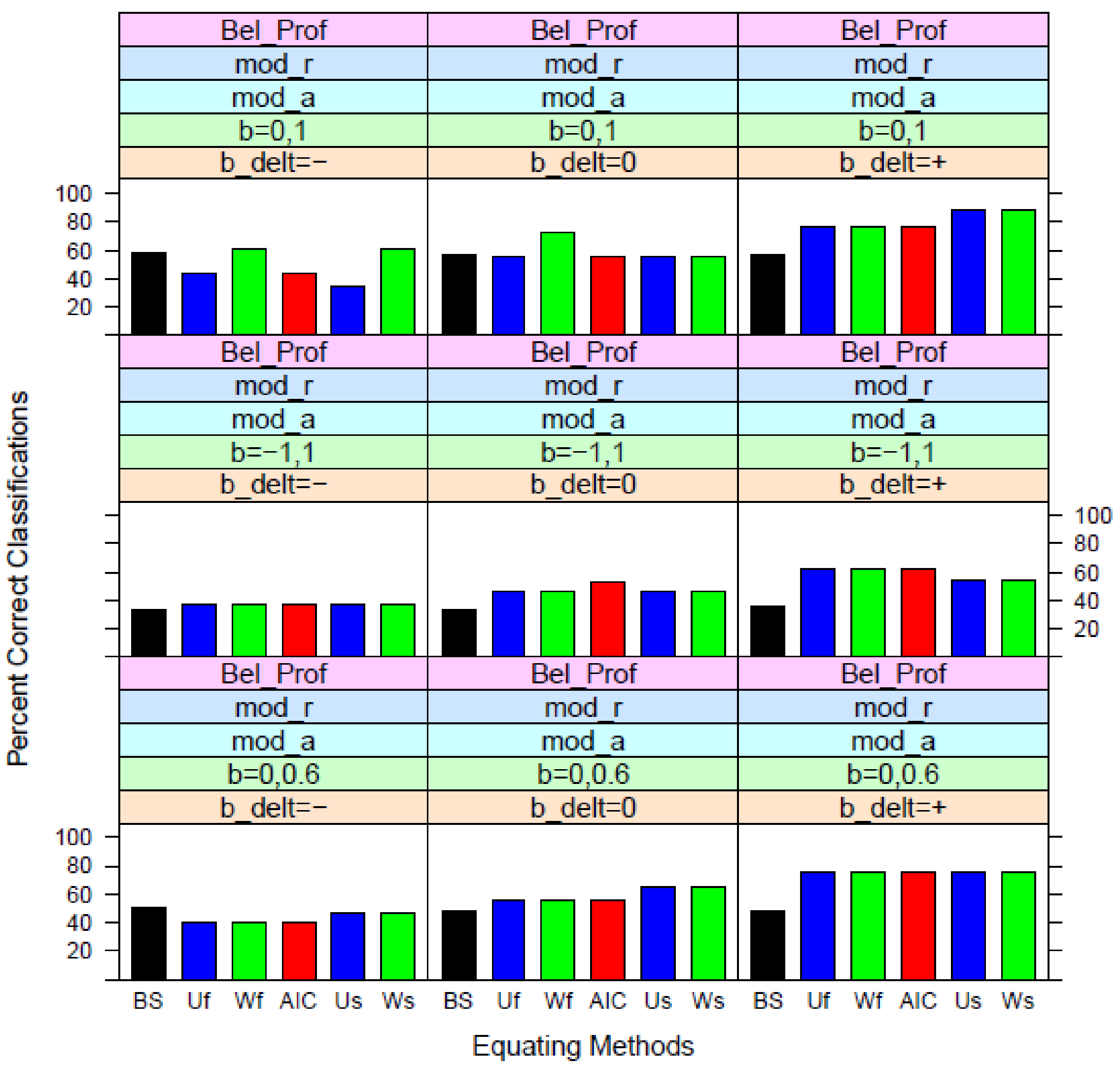


Figure 40. Modified box plot showing bias by equating method. (B = Base form, U = unweighted mean equating, W = weighted mean equating, A = Anchor Item Calibration, U_s = stabilized unweighted mean equating with stabilization and W_s = stabilized weighted mean equating with stabilization)

APPENDIX D

BAR CHARTS SHOWING PERCENT CLASSIFICATION BY EQUATING METHOD



Percent Correct Classification by Equating Method for 50_10 Test Design

Figure 41. Bar charts showing percent classifications by equating method. (B = Base form, U = unweighted mean equating, W = weighted mean equating, A = Anchor Item Calibration, U_s = stabilized unweighted mean equating with stabilization and W_s = stabilized weighted mean equating with stabilization)

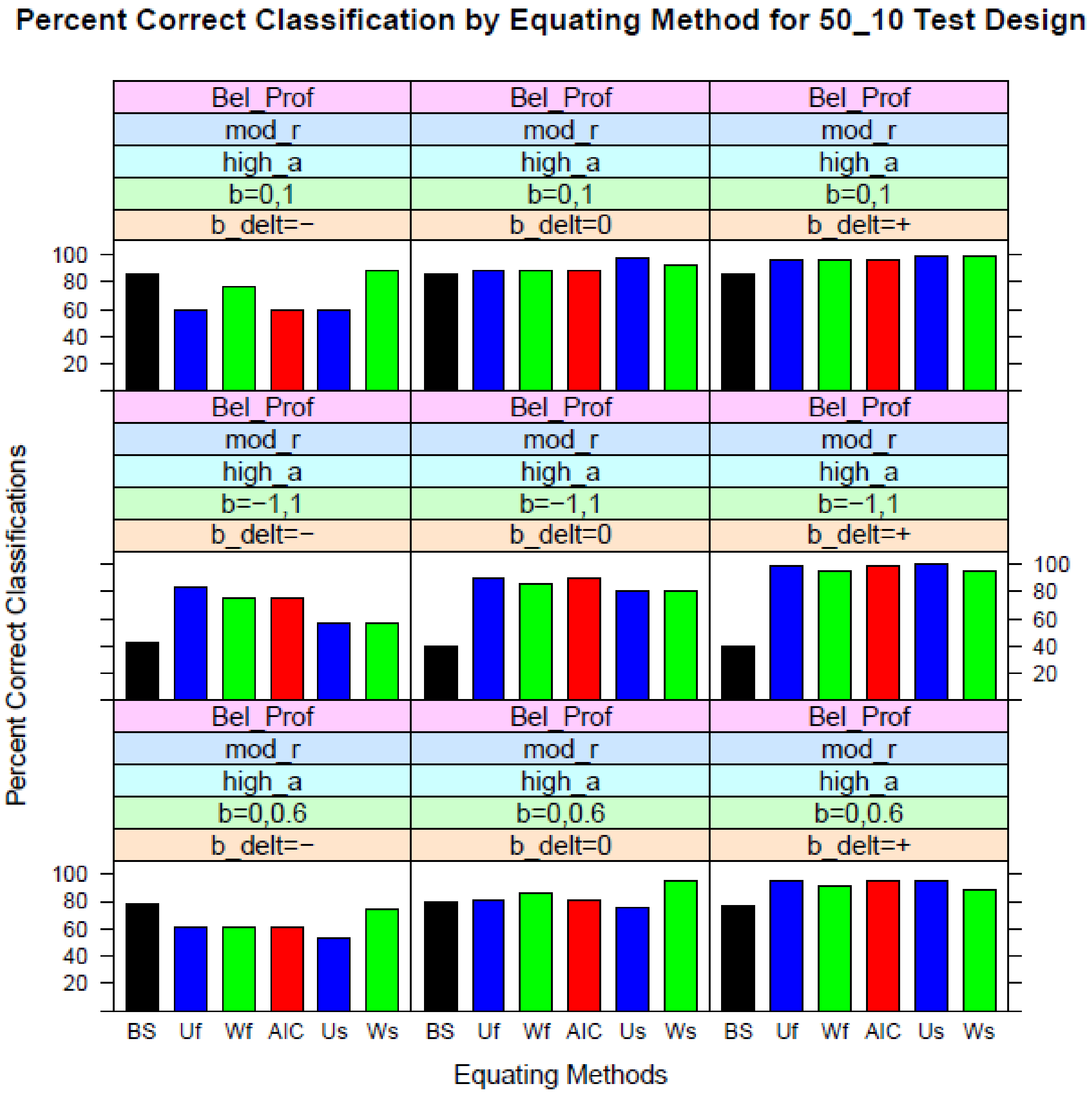


Figure 42. Bar charts showing percent classifications by equating method. (B = Base form, U = unweighted mean equating, W = weighted mean equating, A = Anchor Item Calibration, U_s = stabilized unweighted mean equating with stabilization and W_s = stabilized weighted mean equating with stabilization)

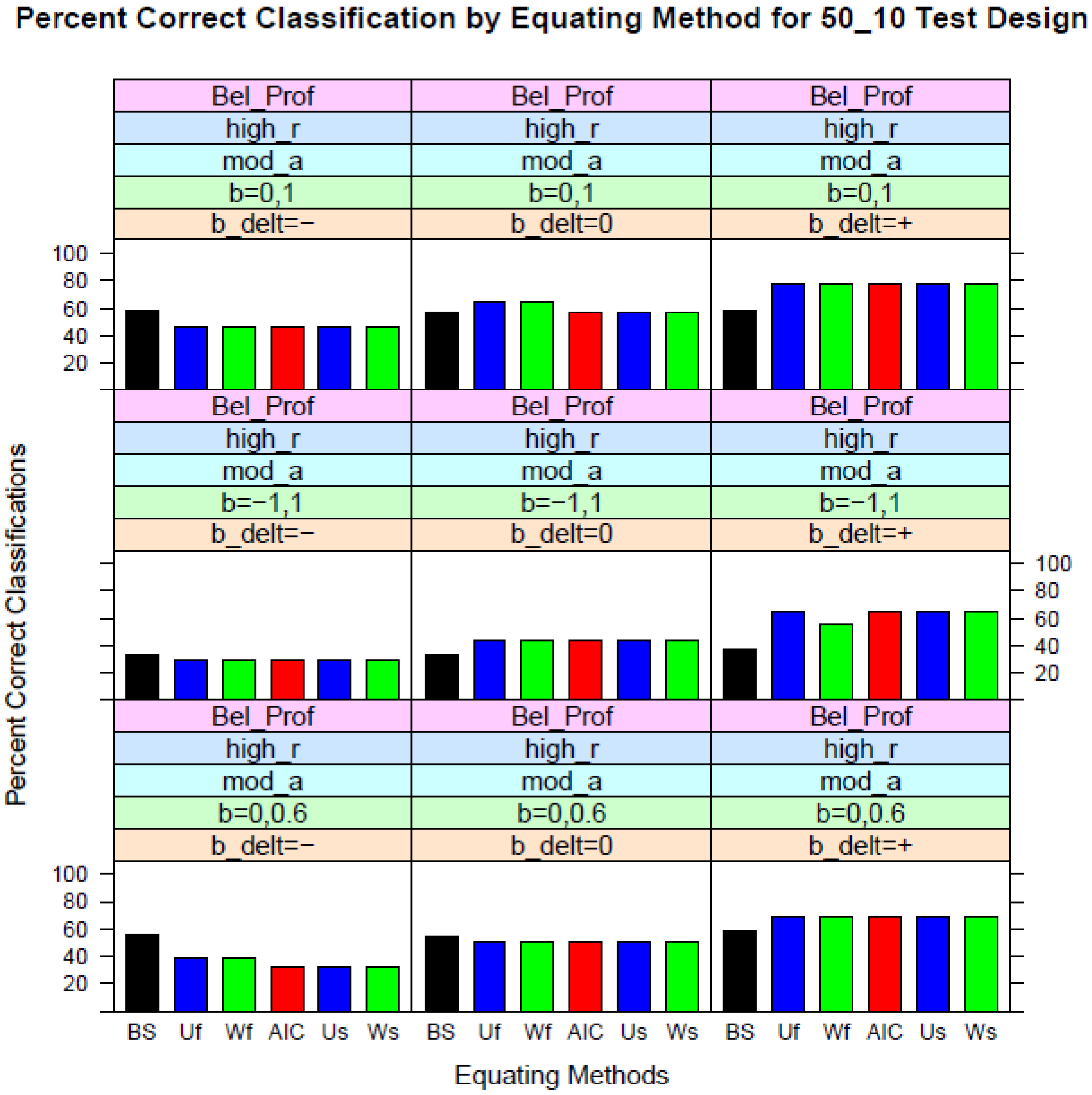


Figure 43. Bar charts showing percent classifications by equating method. (B = Base form, U = unweighted mean equating, W = weighted mean equating, A = Anchor Item Calibration, U_s = stabilized unweighted mean equating with stabilization and W_s = stabilized weighted mean equating with stabilization)

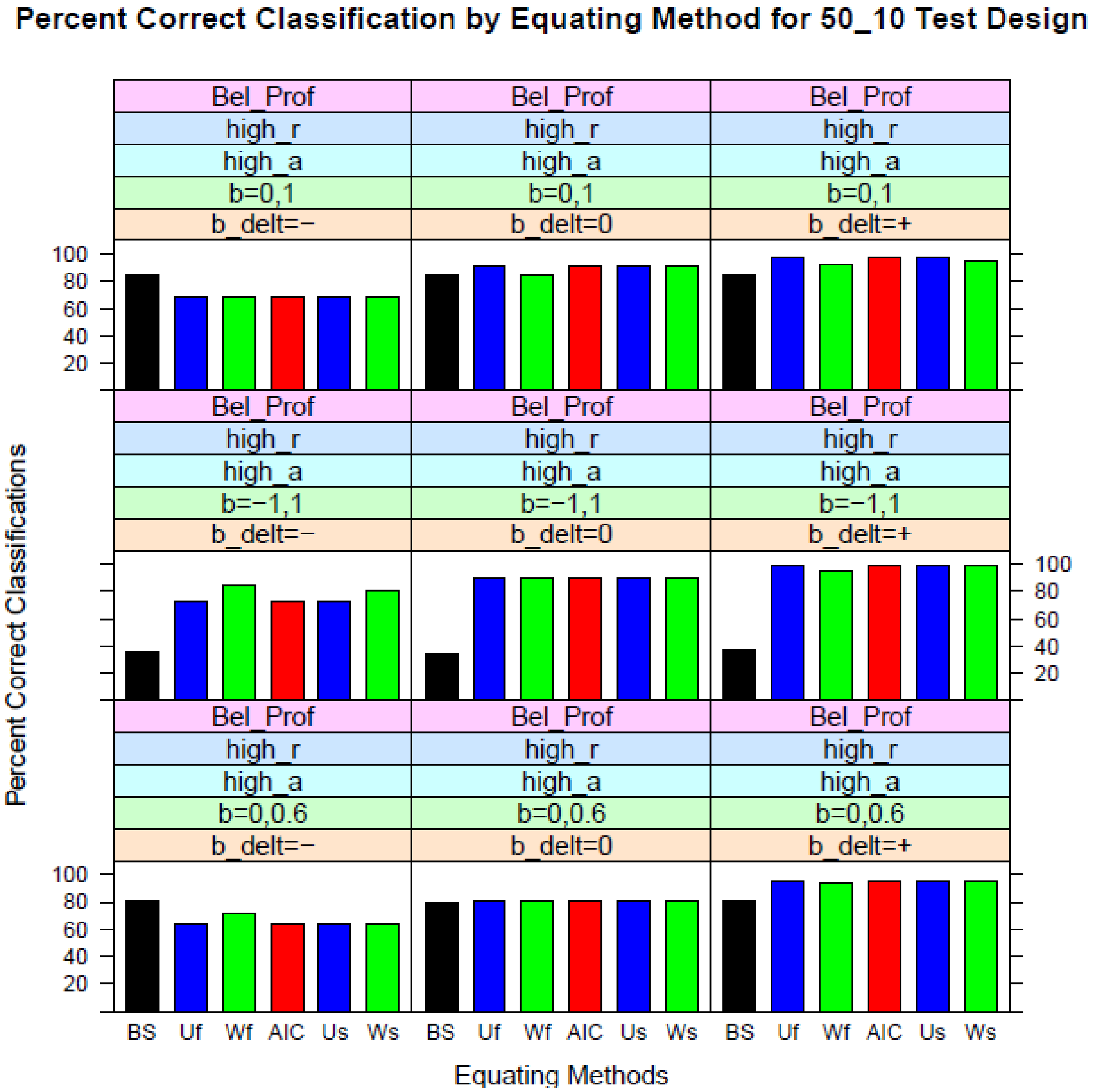


Figure 44. Bar charts showing percent classifications by equating method. (B = Base form, U = unweighted mean equating, W = weighted mean equating, A = Anchor Item Calibration, U_s = stabilized unweighted mean equating with stabilization and W_s = stabilized weighted mean equating with stabilization)

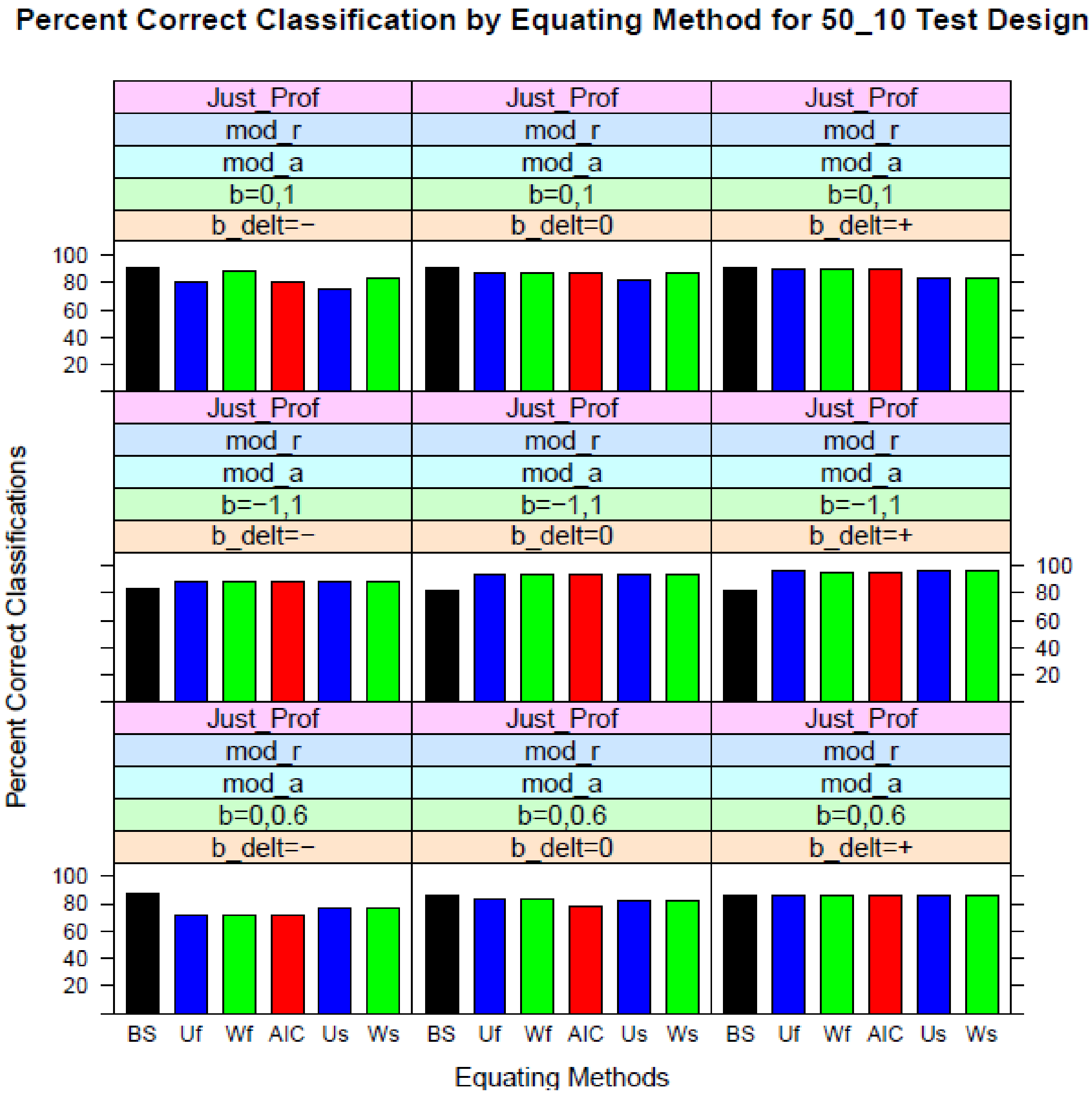


Figure 45. Bar charts showing percent classifications by equating method. (B = Base form, U = unweighted mean equating, W = weighted mean equating, A = Anchor Item Calibration, U_s = stabilized unweighted mean equating with stabilization and W_s = stabilized weighted mean equating with stabilization)

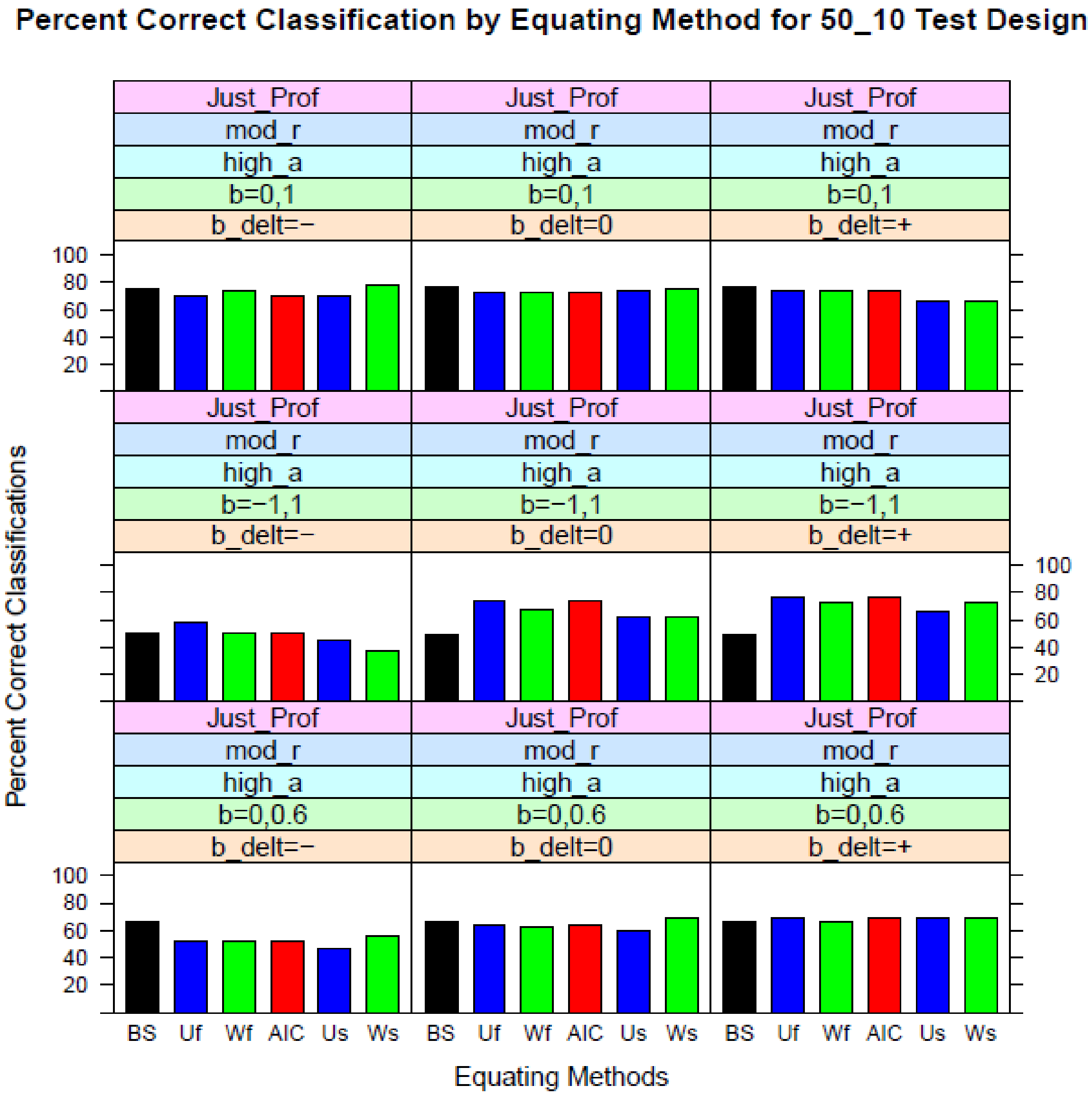


Figure 46. Bar charts showing percent classifications by equating method. (B = Base form, U = unweighted mean equating, W = weighted mean equating, A = Anchor Item Calibration, U_s = stabilized unweighted mean equating with stabilization and W_s = stabilized weighted mean equating with stabilization)

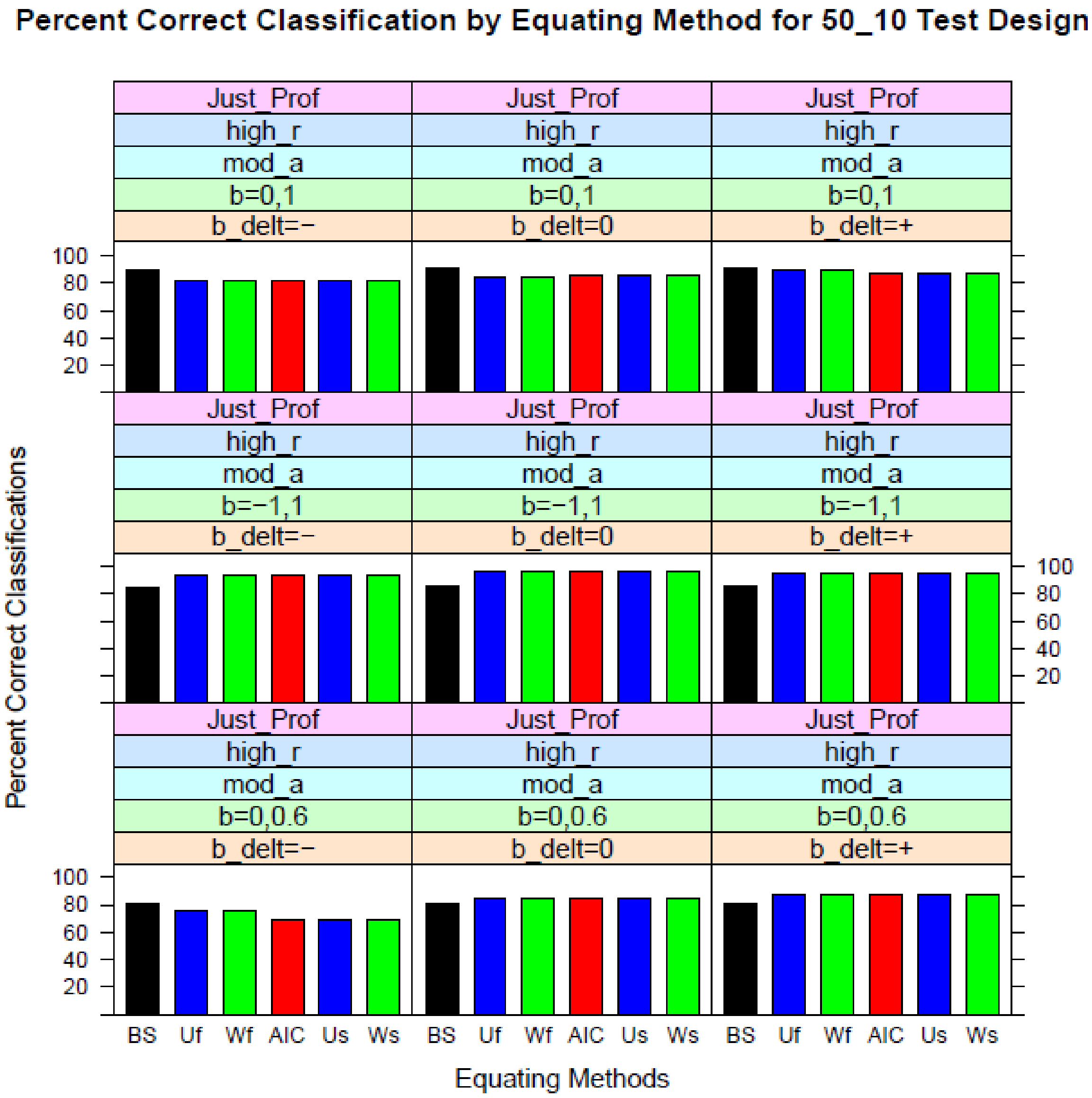


Figure 47. Bar charts showing percent classifications by equating method. (B = Base form, U = unweighted mean equating, W = weighted mean equating, A = Anchor Item Calibration, U_s = stabilized unweighted mean equating with stabilization and W_s = stabilized weighted mean equating with stabilization)

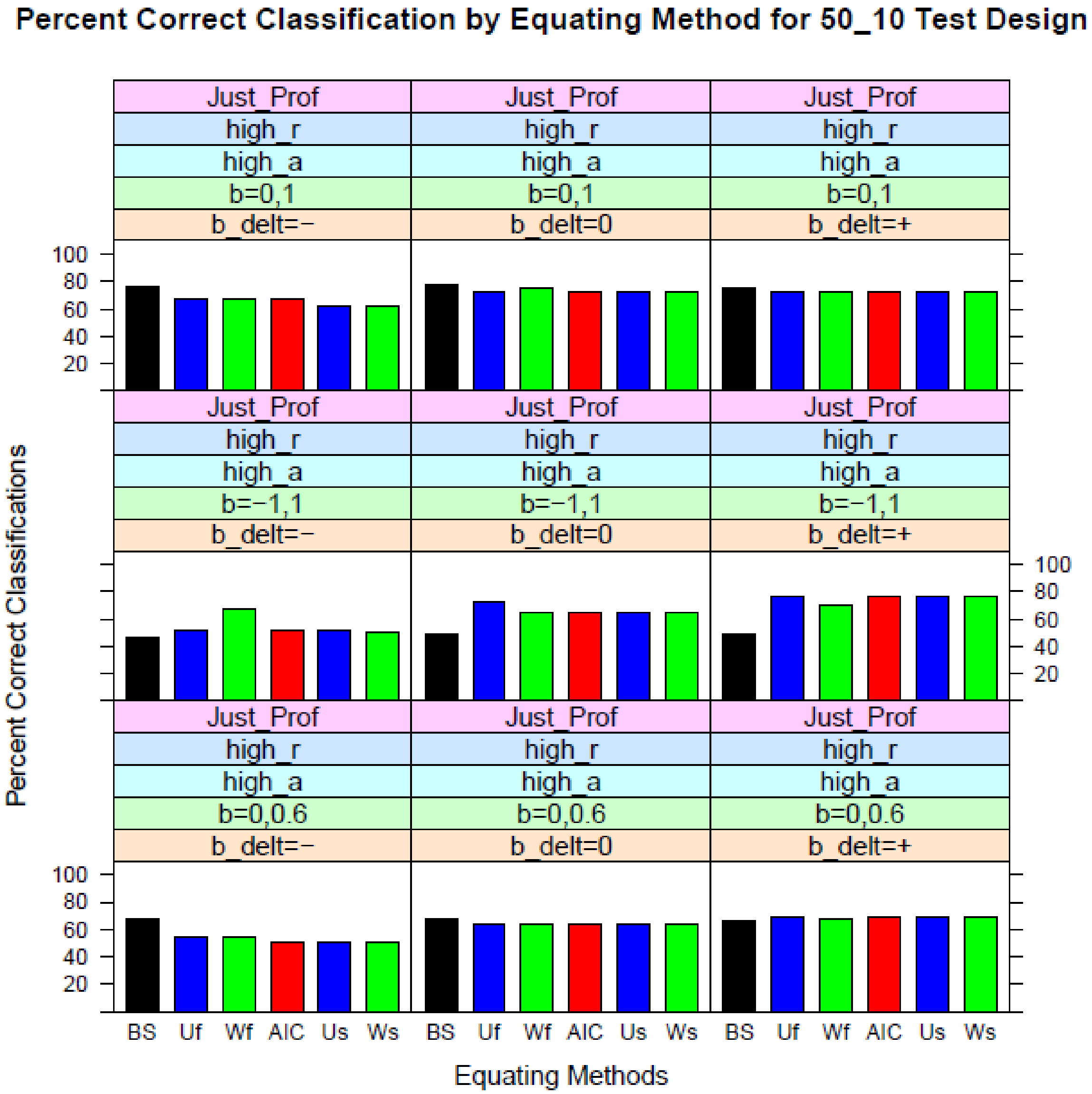


Figure 48. Bar charts showing percent classifications by equating method. (B = Base form, U = unweighted mean equating, W = weighted mean equating, A = Anchor Item Calibration, U_s = stabilized unweighted mean equating with stabilization and W_s = stabilized weighted mean equating with stabilization)

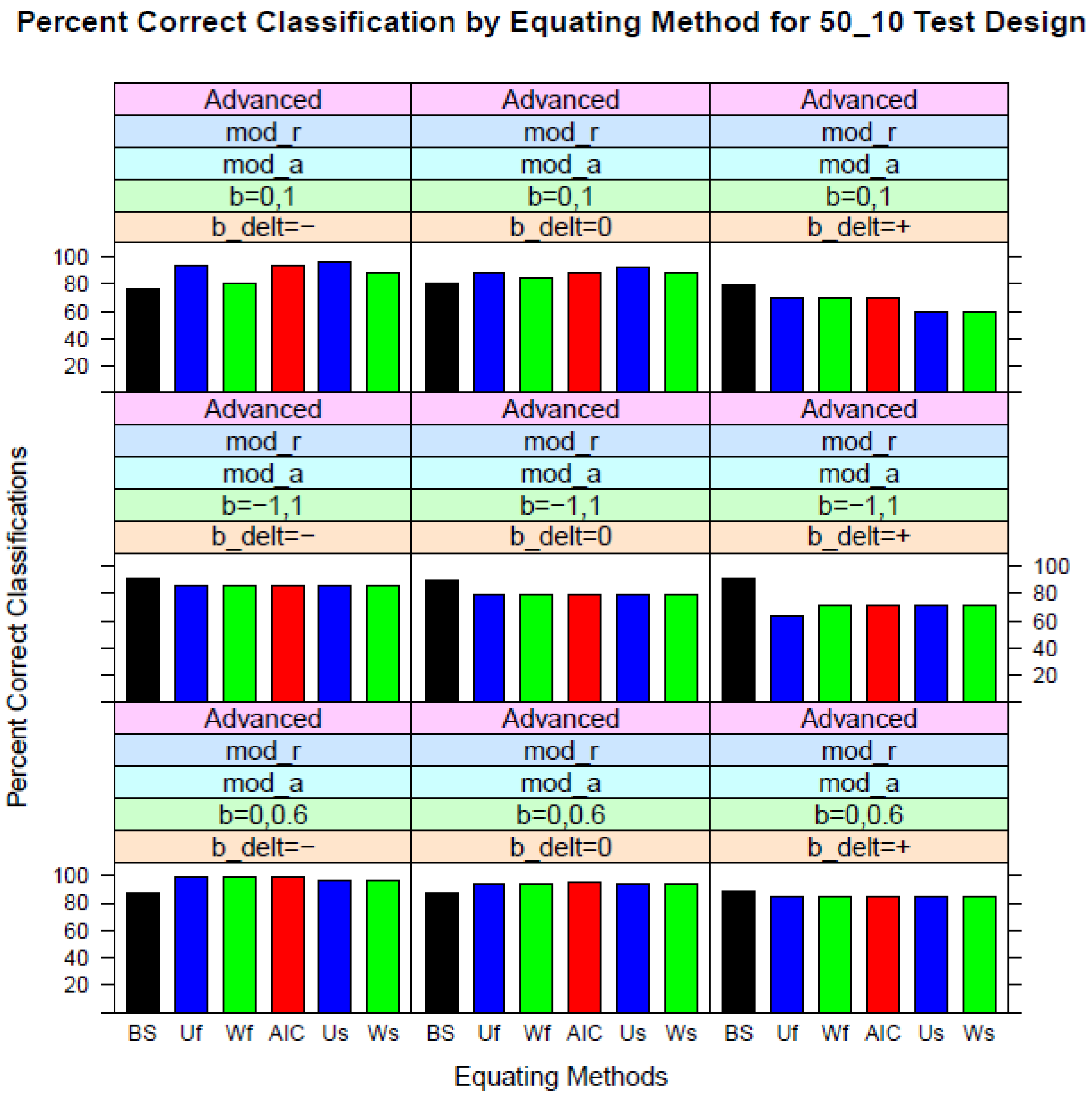


Figure 49. Bar charts showing percent classifications by equating method. (B = Base form, U = unweighted mean equating, W = weighted mean equating, A = Anchor Item Calibration, U_s = stabilized unweighted mean equating with stabilization and W_s = stabilized weighted mean equating with stabilization)

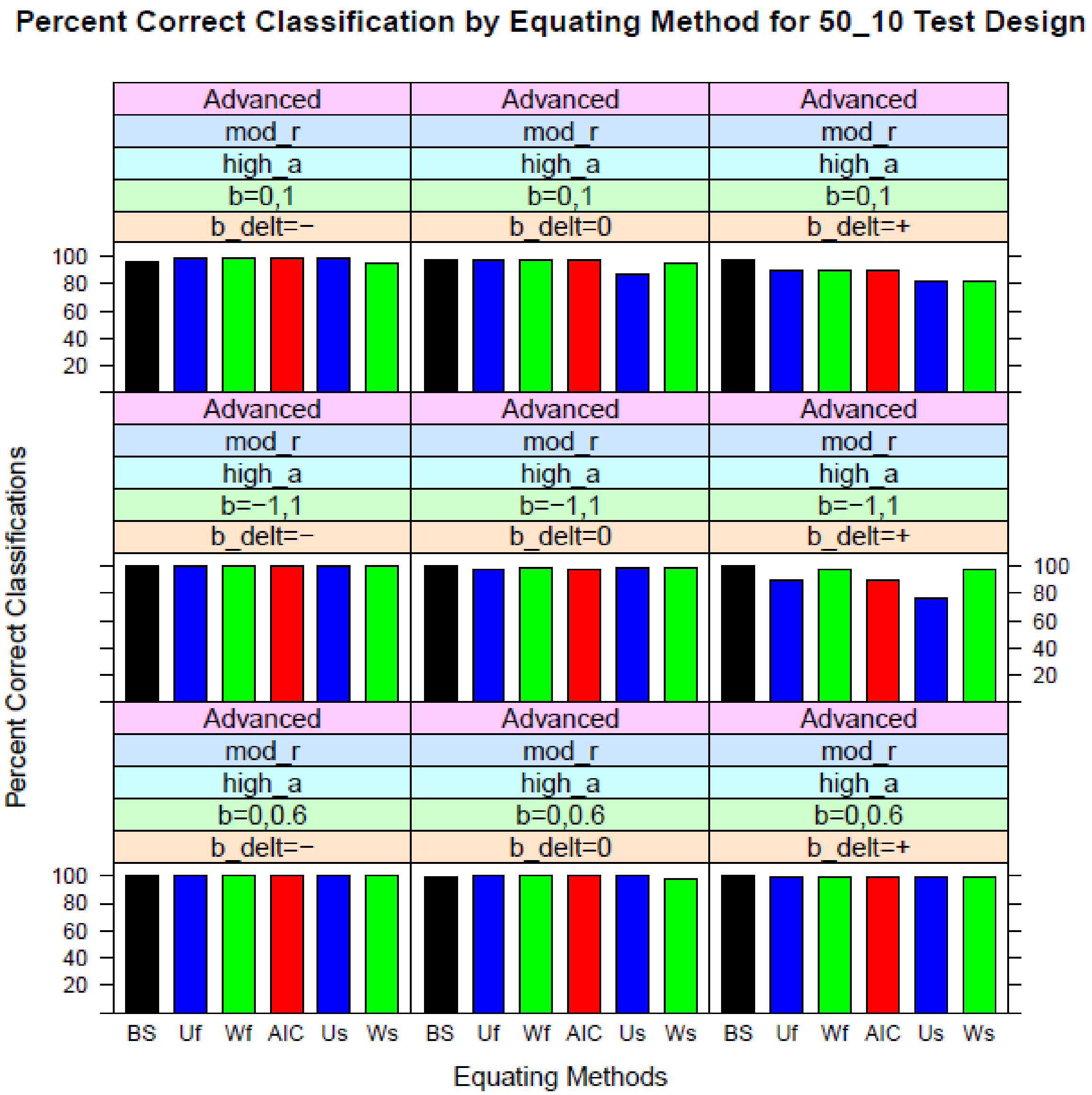


Figure 50. Bar charts showing percent classifications by equating method. (B = Base form, U = unweighted mean equating, W = weighted mean equating, A = Anchor Item Calibration, U_s = stabilized unweighted mean equating with stabilization and W_s = stabilized weighted mean equating with stabilization)

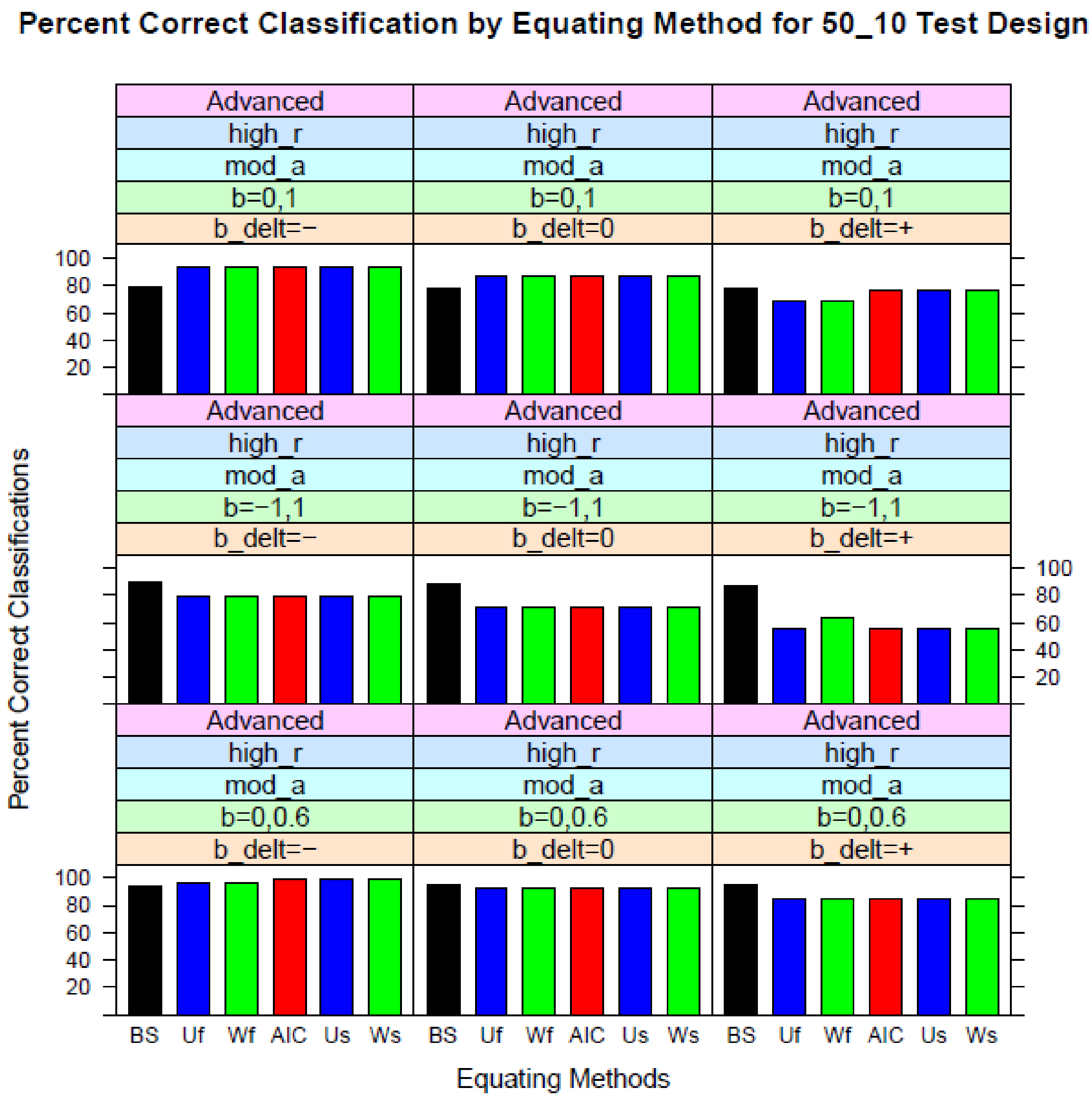


Figure 51. Bar charts showing percent classifications by equating method. (B = Base form, U = unweighted mean equating, W = weighted mean equating, A = Anchor Item Calibration, U_s = stabilized unweighted mean equating with stabilization and W_s = stabilized weighted mean equating with stabilization)

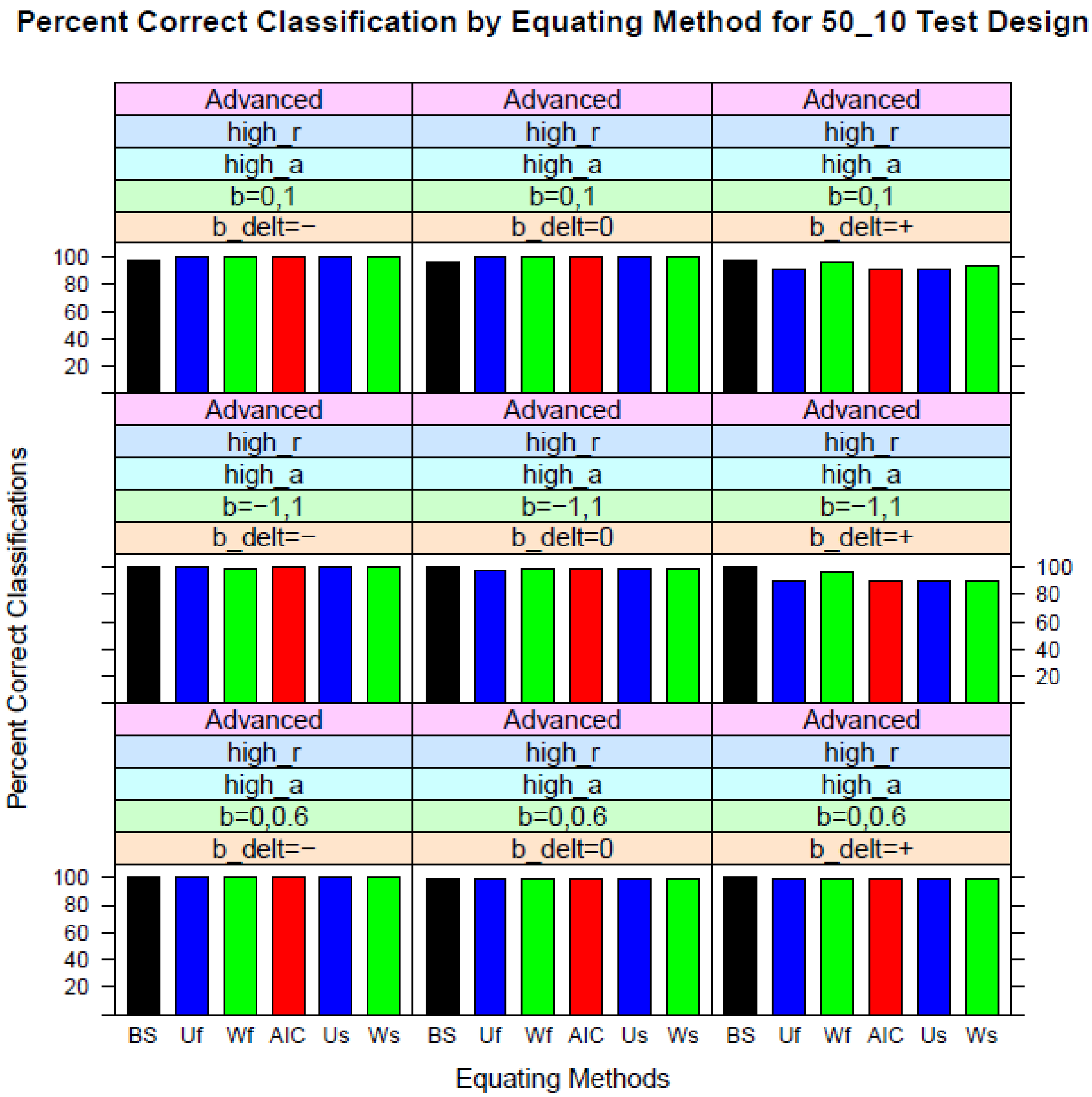


Figure 52. Bar charts showing percent classifications by equating method. (B = Base form, U = unweighted mean equating, W = weighted mean equating, A = Anchor Item Calibration, U_s = stabilized unweighted mean equating with stabilization and W_s = stabilized weighted mean equating with stabilization)

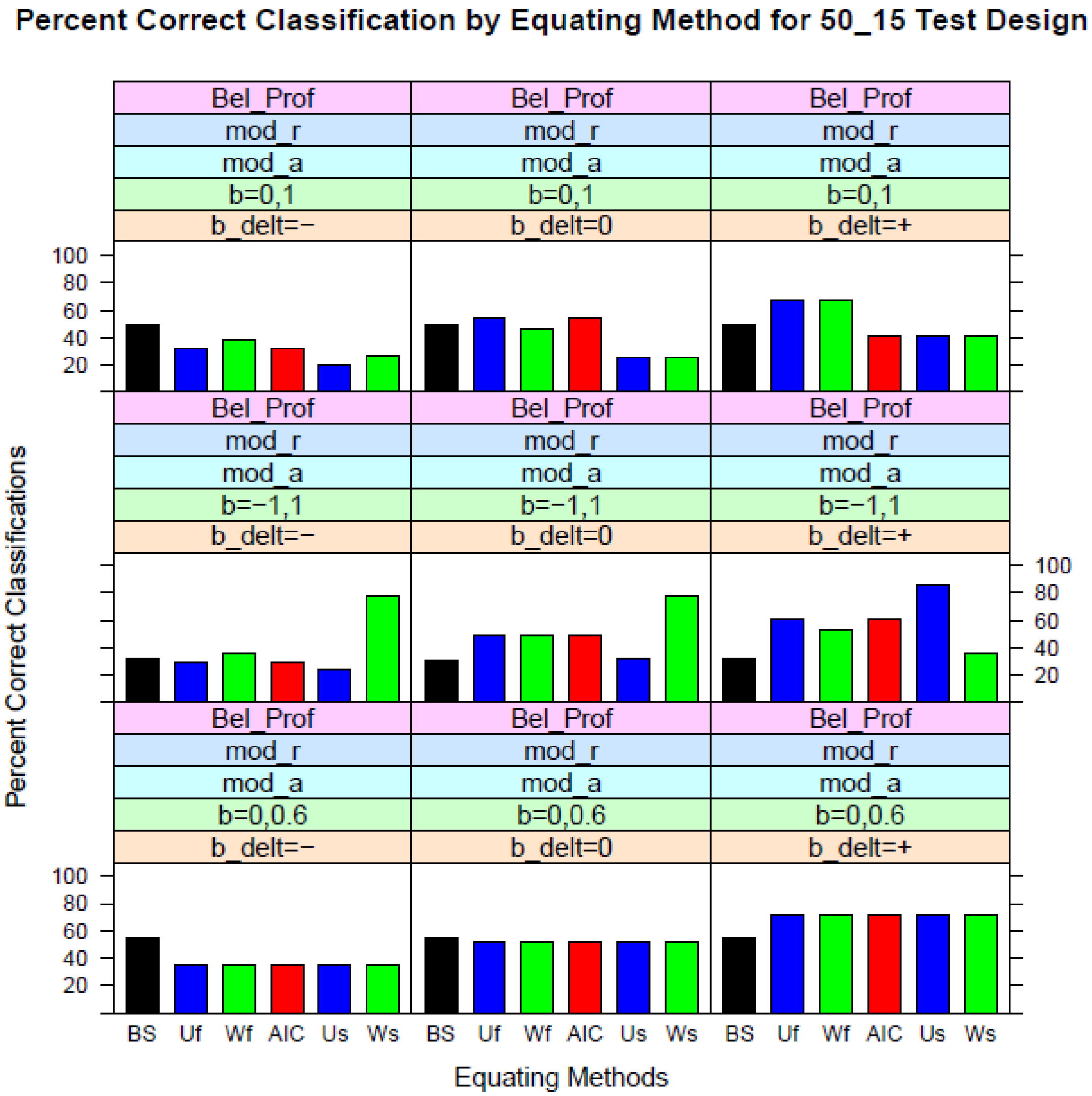


Figure 53. Bar charts showing percent classifications by equating method. (B = Base form, U = unweighted mean equating, W = weighted mean equating, A = Anchor Item Calibration, U_s = stabilized unweighted mean equating with stabilization and W_s = stabilized weighted mean equating with stabilization)

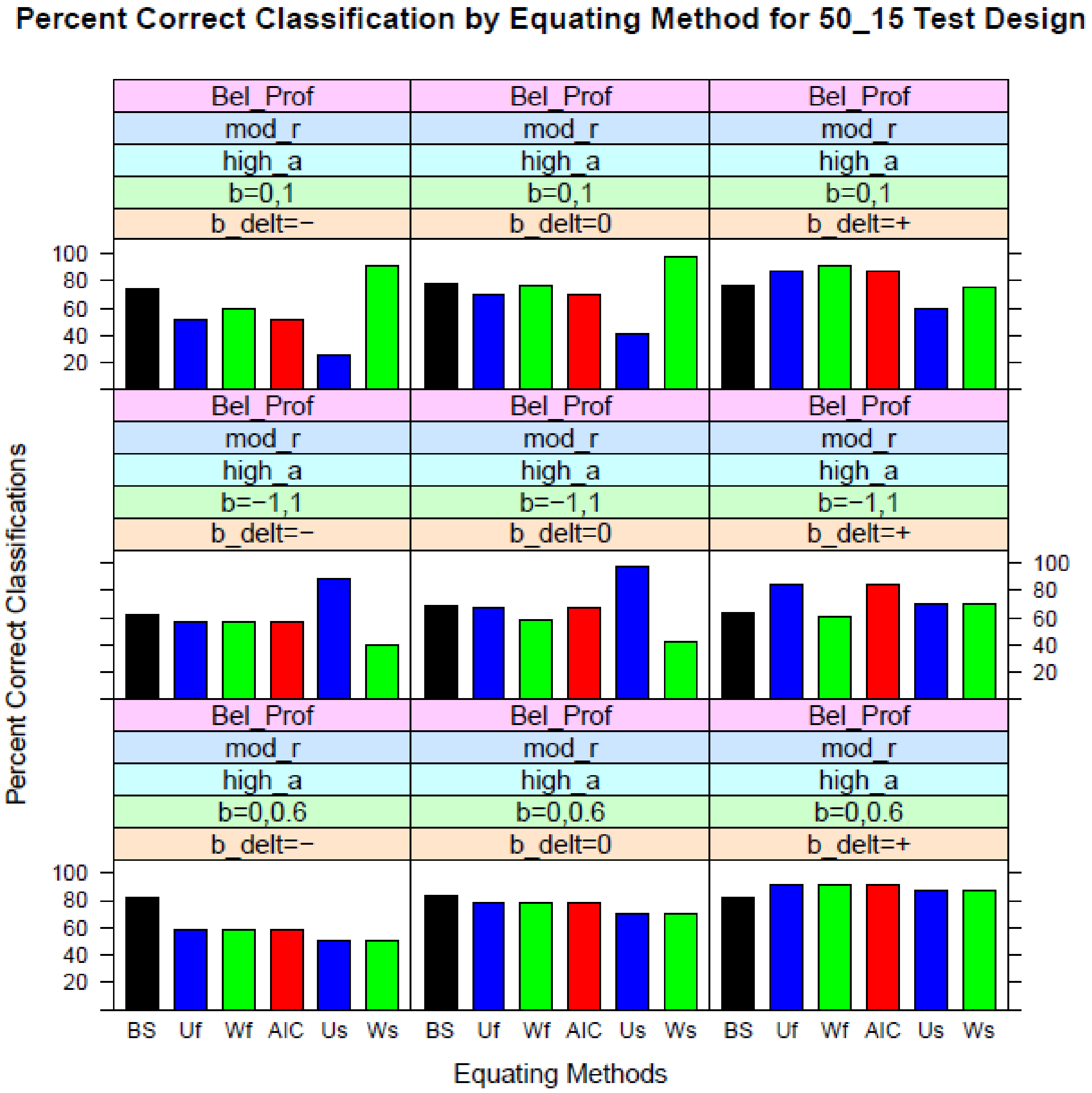


Figure 54. Bar charts showing percent classifications by equating method. (B = Base form, U = unweighted mean equating, W = weighted mean equating, A = Anchor Item Calibration, U_s = stabilized unweighted mean equating with stabilization and W_s = stabilized weighted mean equating with stabilization)

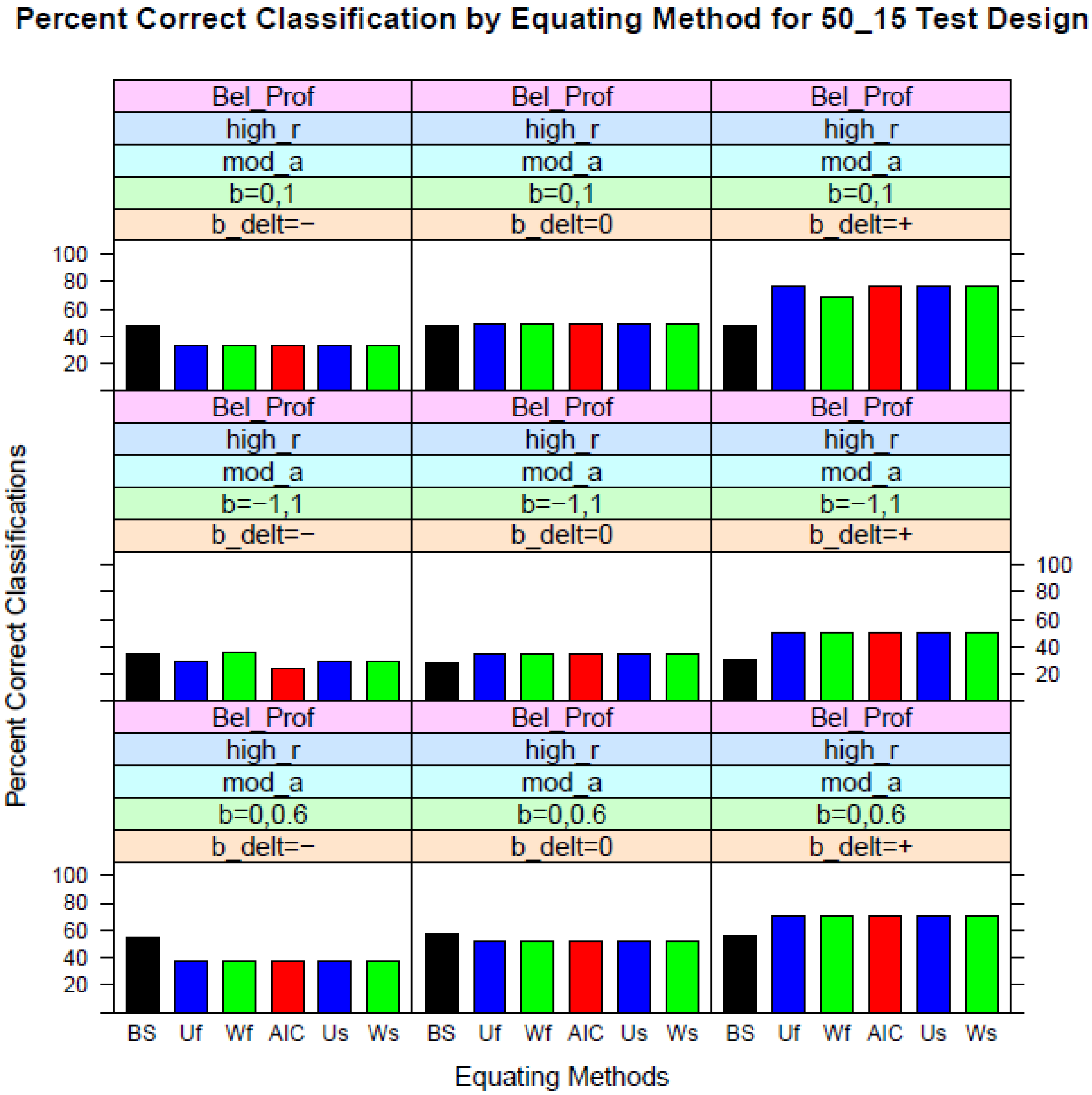


Figure 55. Bar charts showing percent classifications by equating method. (B = Base form, U = unweighted mean equating, W = weighted mean equating, A = Anchor Item Calibration, U_s = stabilized unweighted mean equating with stabilization and W_s = stabilized weighted mean equating with stabilization)

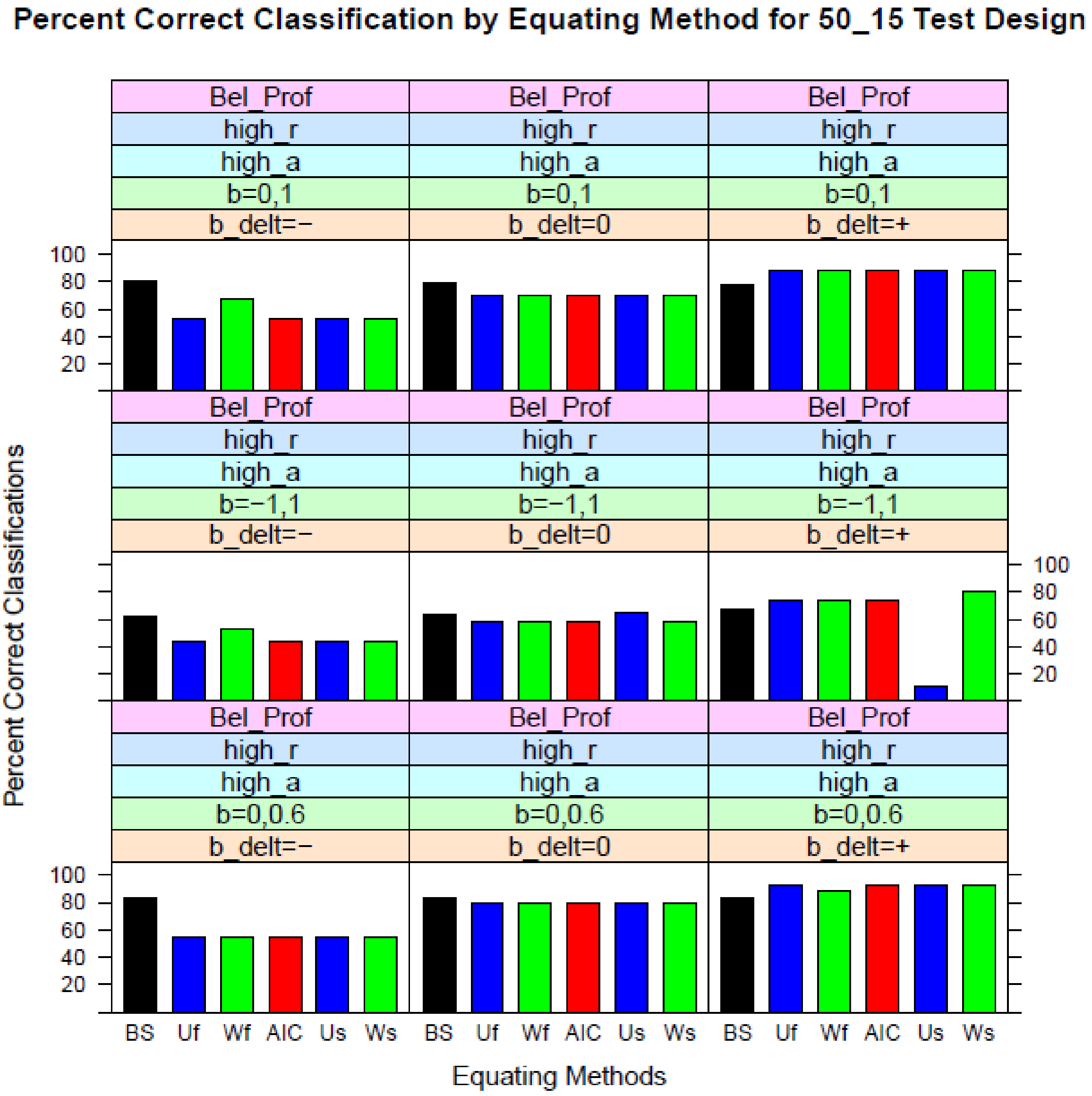


Figure 56. Bar charts showing percent classifications by equating method. (B = Base form, U = unweighted mean equating, W = weighted mean equating, A = Anchor Item Calibration, U_s = stabilized unweighted mean equating with stabilization and W_s = stabilized weighted mean equating with stabilization)

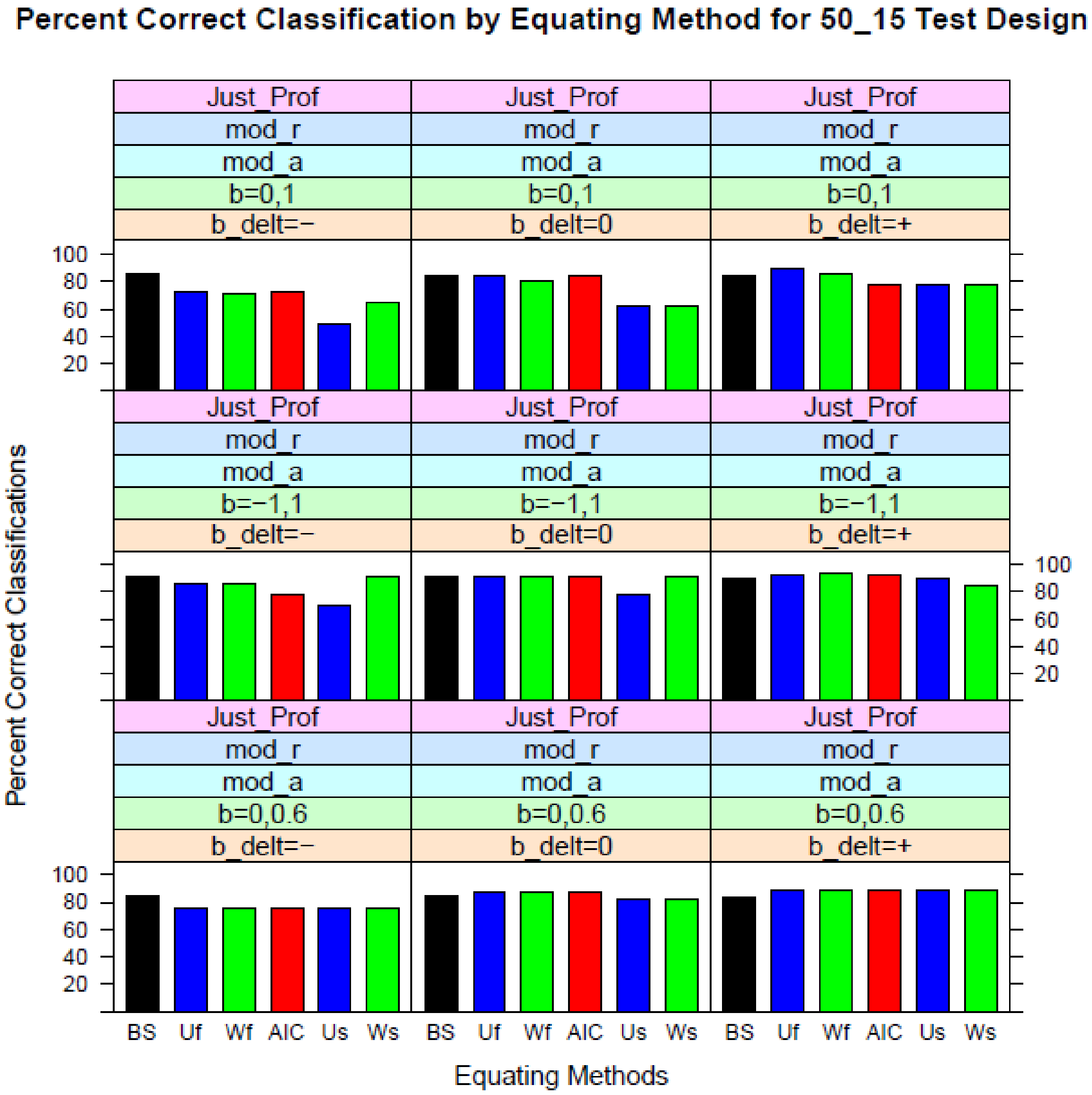


Figure 57. Bar charts showing percent classifications by equating method. (B = Base form, U = unweighted mean equating, W = weighted mean equating, A = Anchor Item Calibration, U_s = stabilized unweighted mean equating with stabilization and W_s = stabilized weighted mean equating with stabilization)

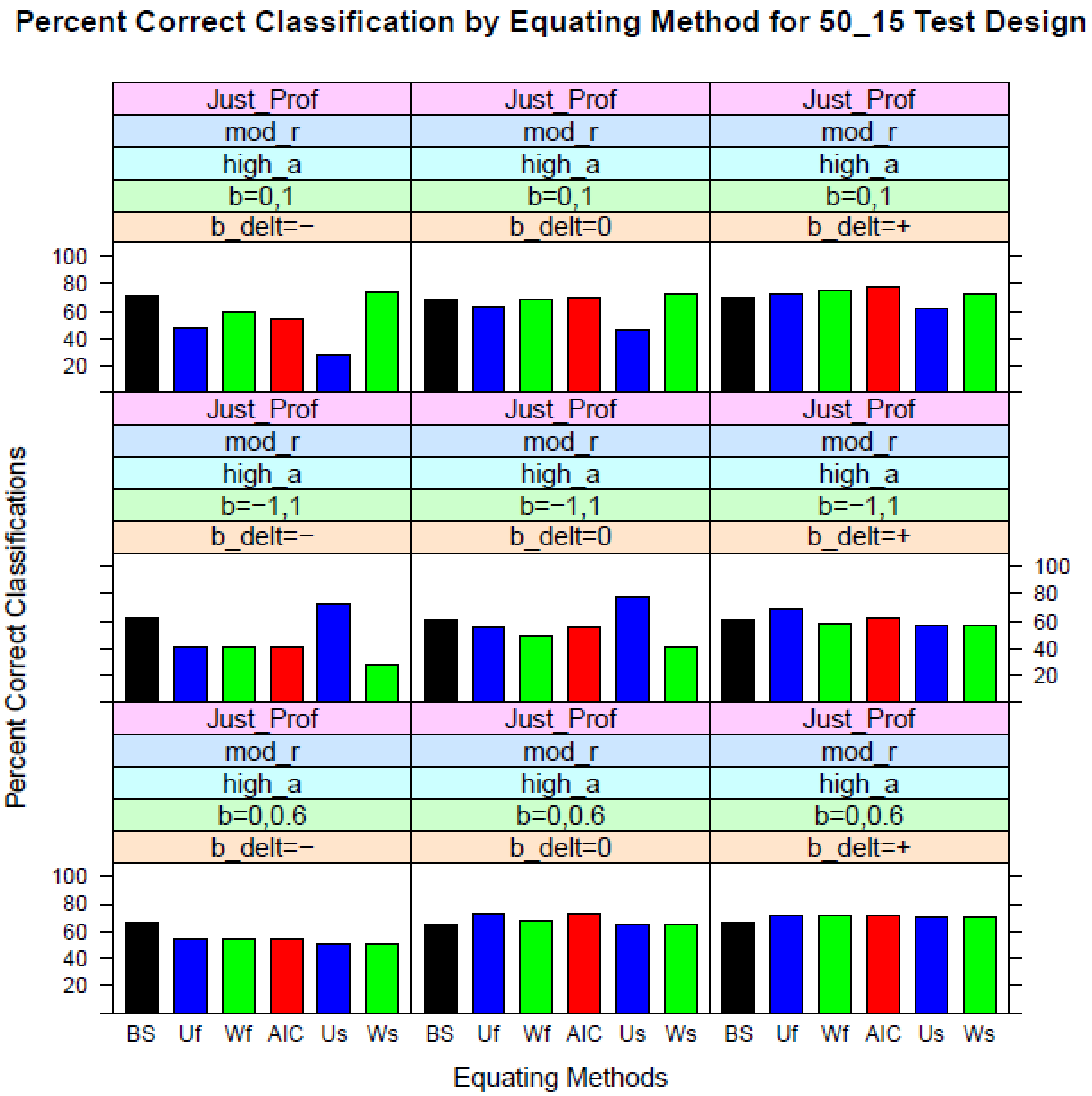


Figure 58. Bar charts showing percent classifications by equating method. (B = Base form, U = unweighted mean equating, W = weighted mean equating, A = Anchor Item Calibration, U_s = stabilized unweighted mean equating with stabilization and W_s = stabilized weighted mean equating with stabilization)

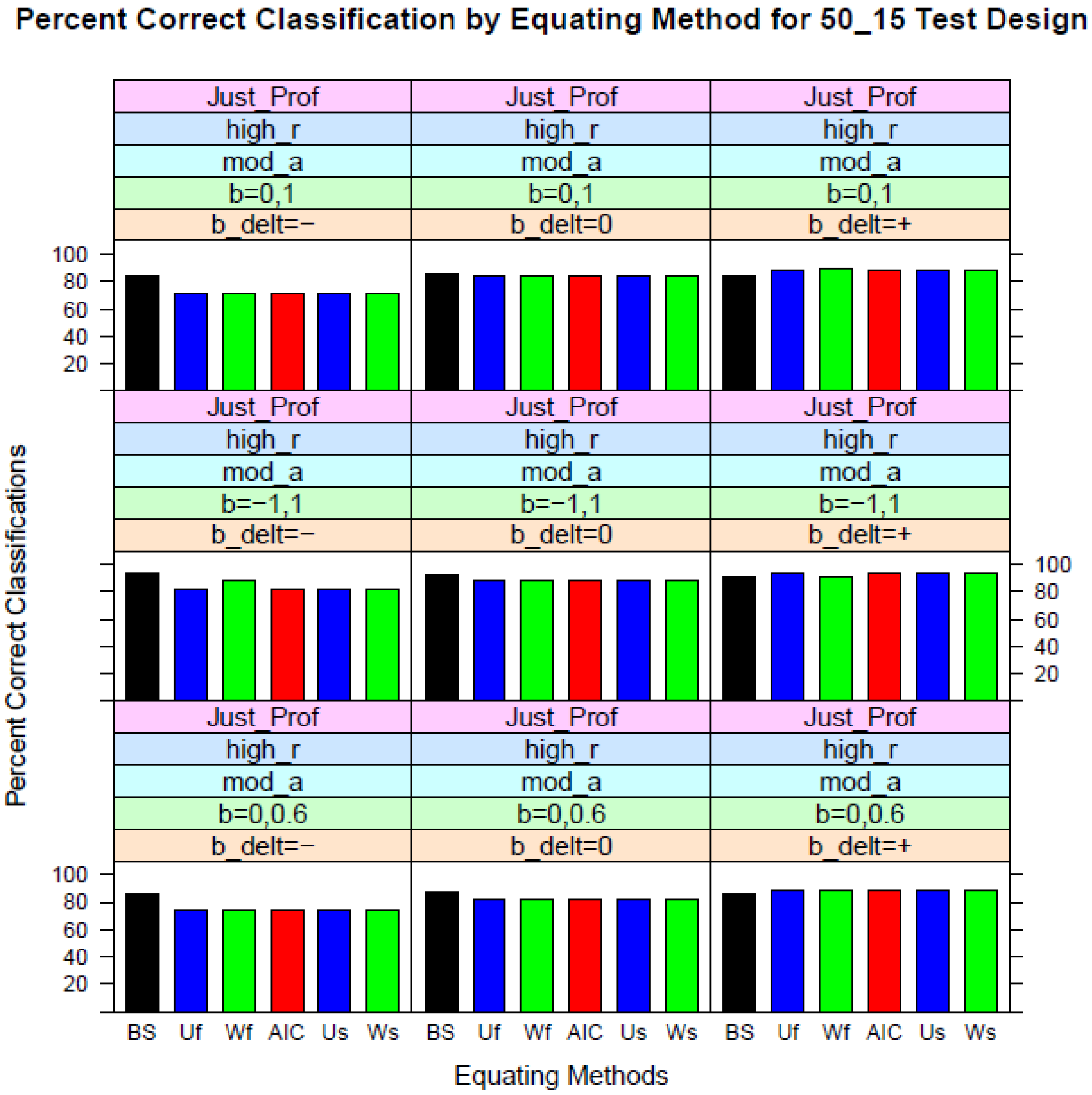


Figure 59. Bar charts showing percent classifications by equating method. (B = Base form, U = unweighted mean equating, W = weighted mean equating, A = Anchor Item Calibration, U_s = stabilized unweighted mean equating with stabilization and W_s = stabilized weighted mean equating with stabilization)

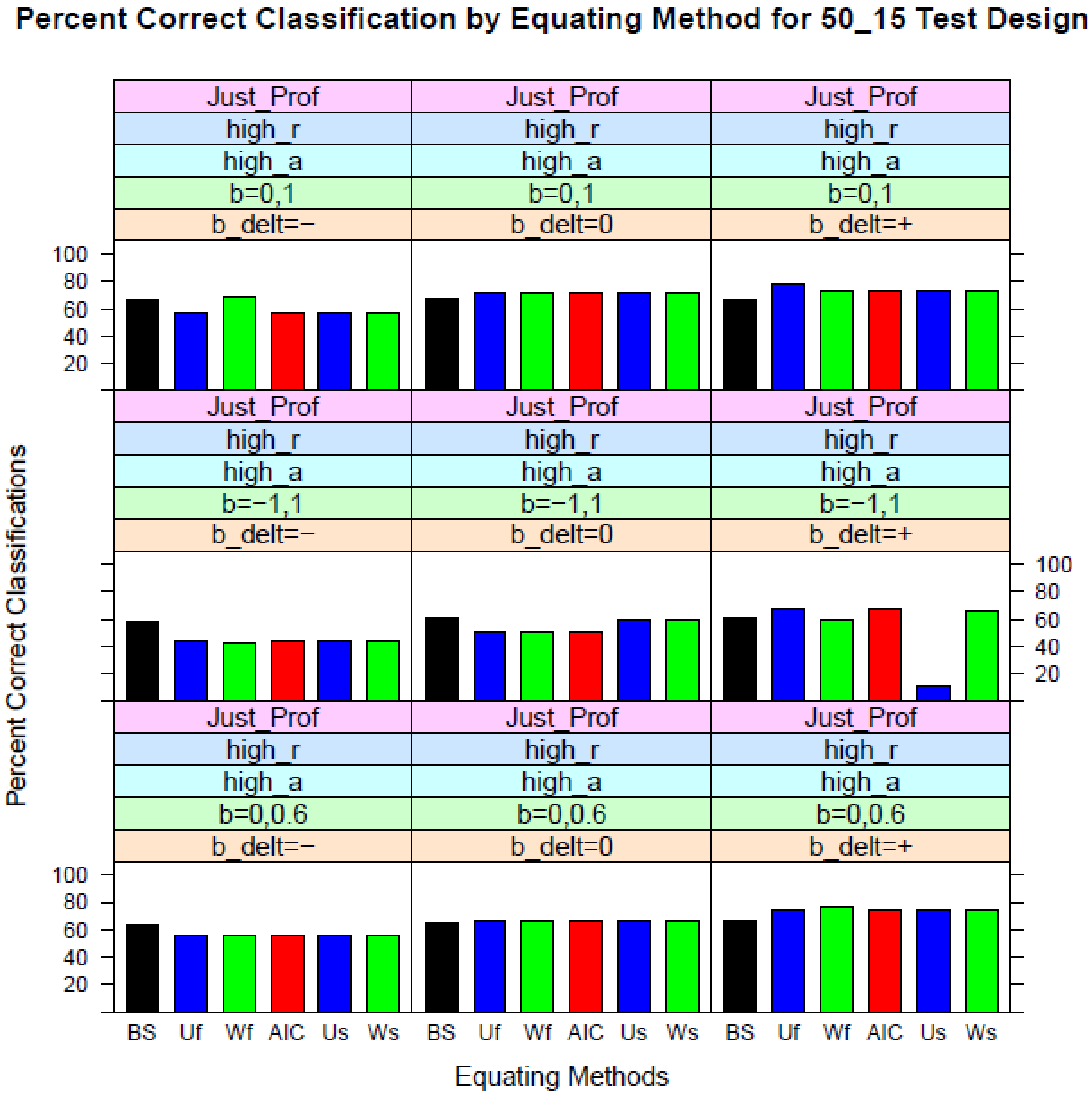


Figure 60. Bar charts showing percent classifications by equating method. (B = Base form, U = unweighted mean equating, W = weighted mean equating, A = Anchor Item Calibration, U_s = stabilized unweighted mean equating with stabilization and W_s = stabilized weighted mean equating with stabilization)

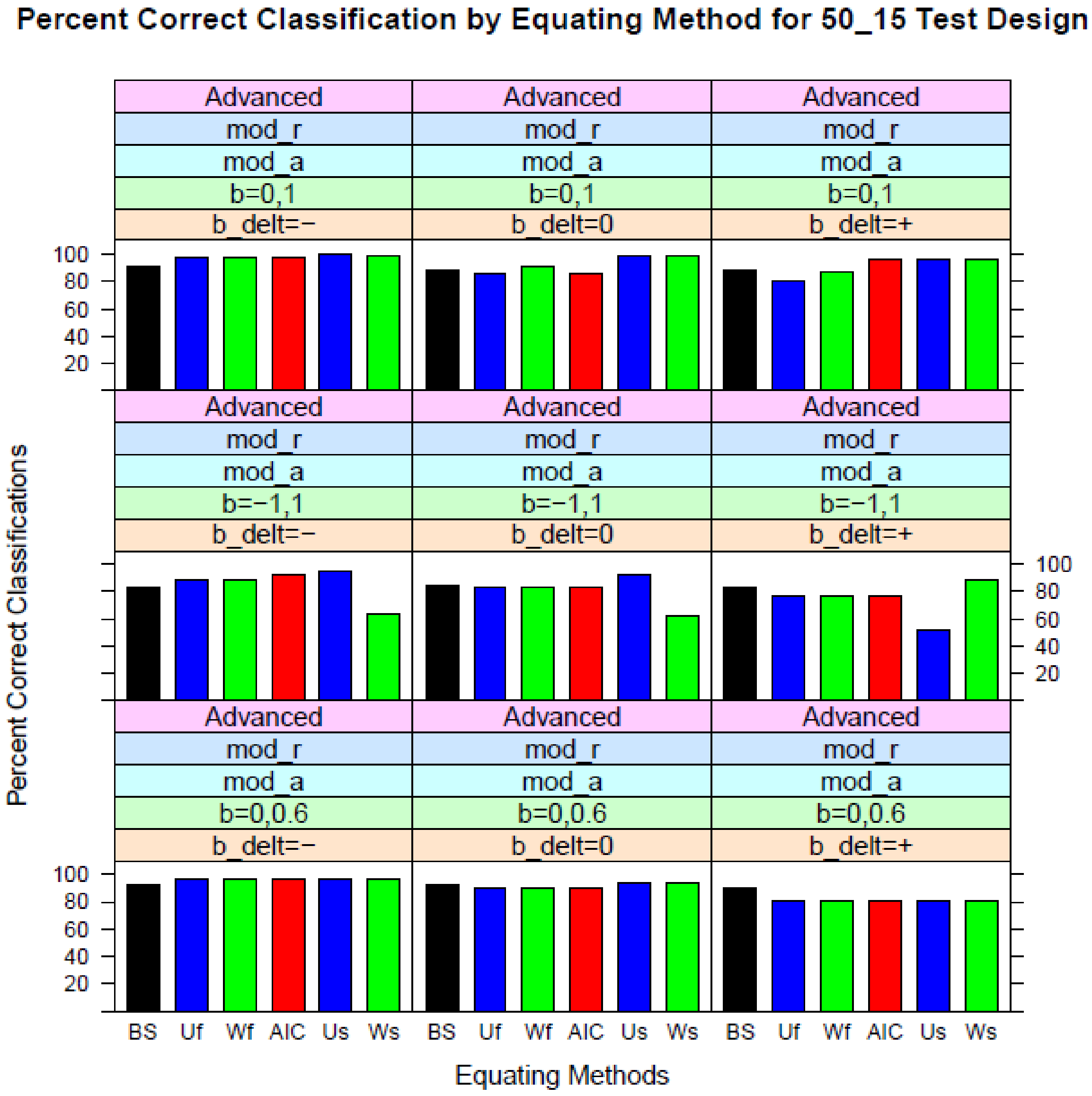


Figure 61. Bar charts showing percent classifications by equating method. (B = Base form, U = unweighted mean equating, W = weighted mean equating, A = Anchor Item Calibration, U_s = stabilized unweighted mean equating with stabilization and W_s = stabilized weighted mean equating with stabilization)

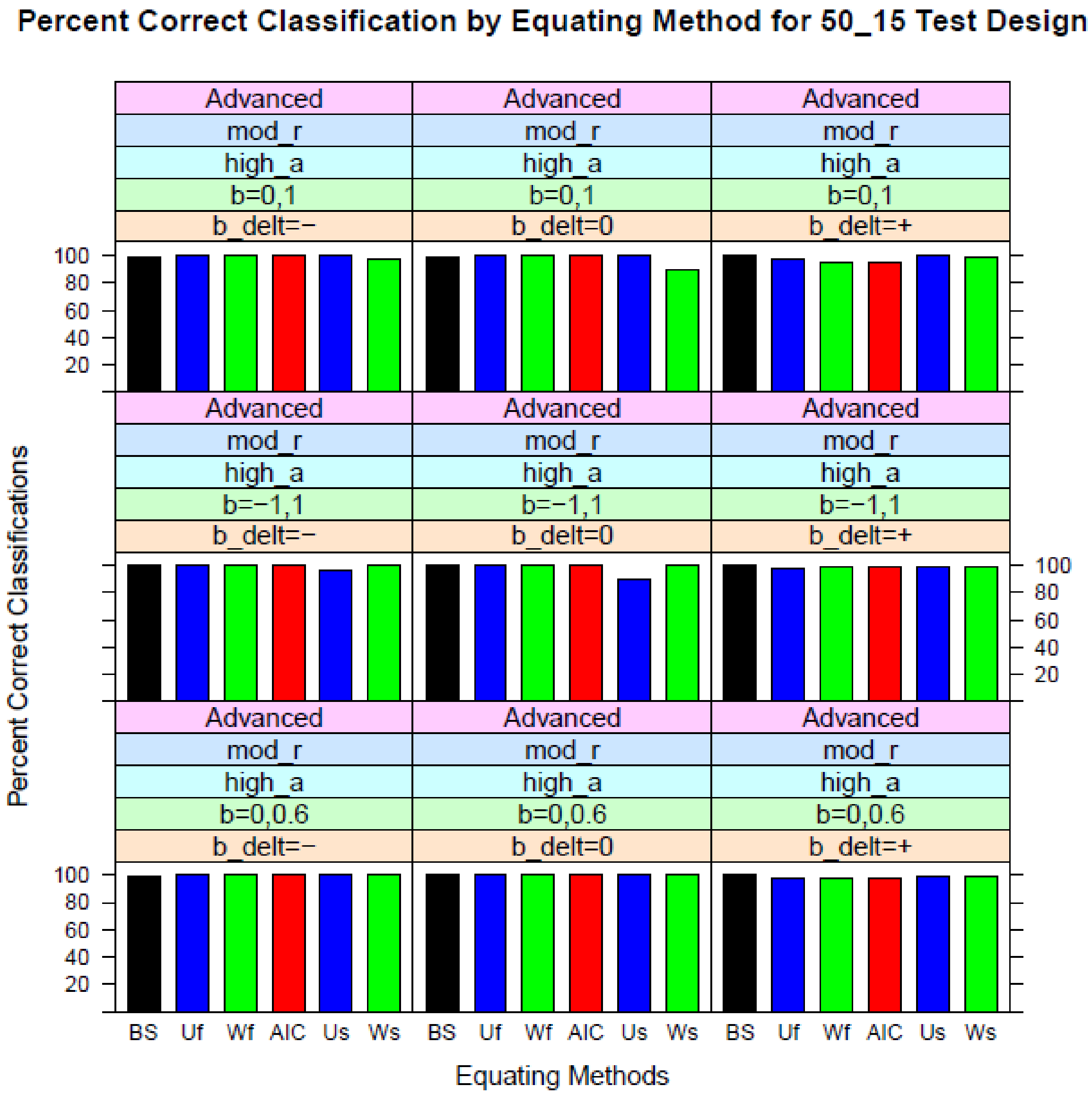


Figure 62. Bar charts showing percent classifications by equating method. (B = Base form, U = unweighted mean equating, W = weighted mean equating, A = Anchor Item Calibration, U_s = stabilized unweighted mean equating with stabilization and W_s = stabilized weighted mean equating with stabilization)

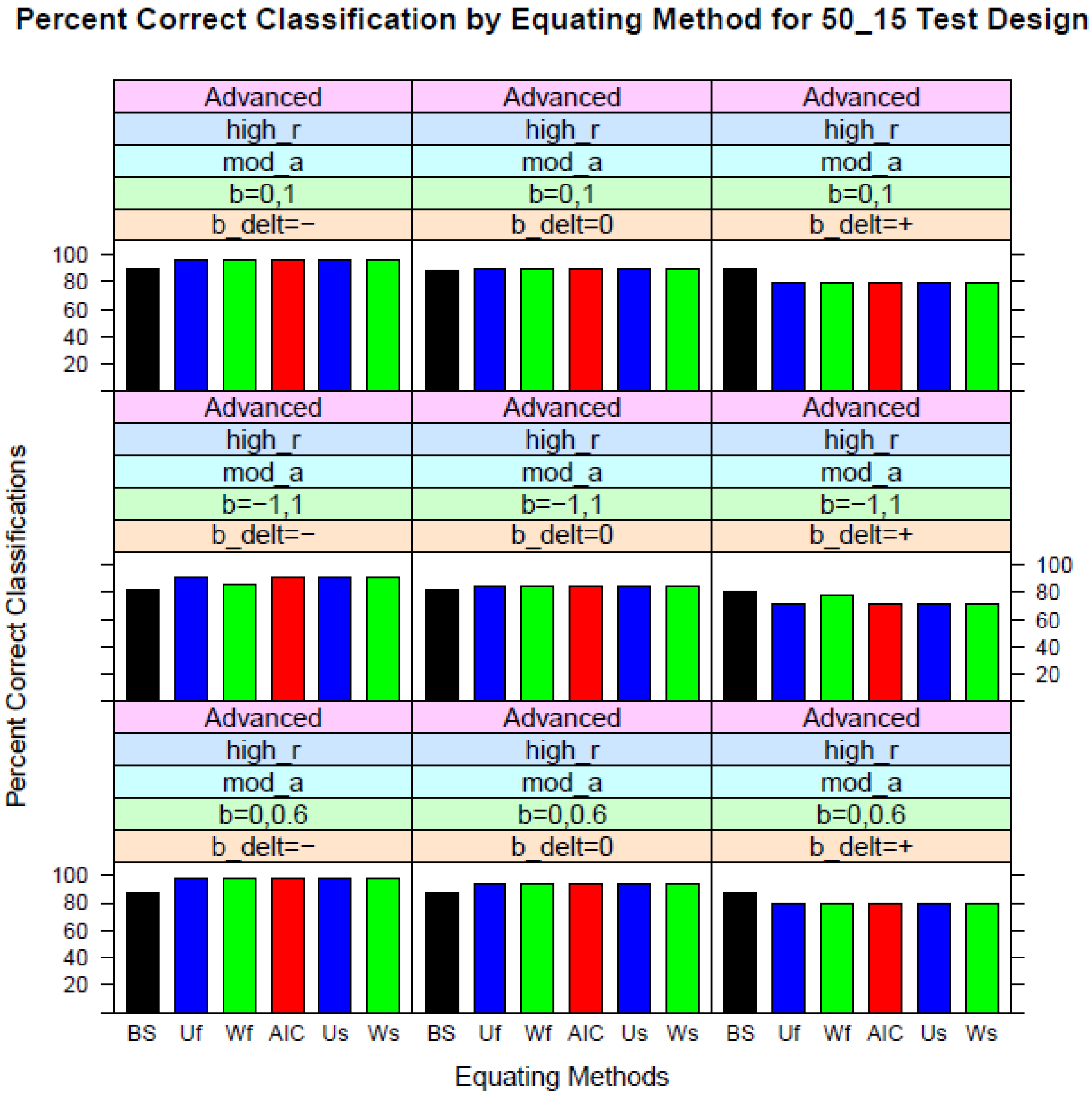


Figure 63. Bar charts showing percent classifications by equating method. (B = Base form, U = unweighted mean equating, W = weighted mean equating, A = Anchor Item Calibration, U_s = stabilized unweighted mean equating with stabilization and W_s = stabilized weighted mean equating with stabilization)

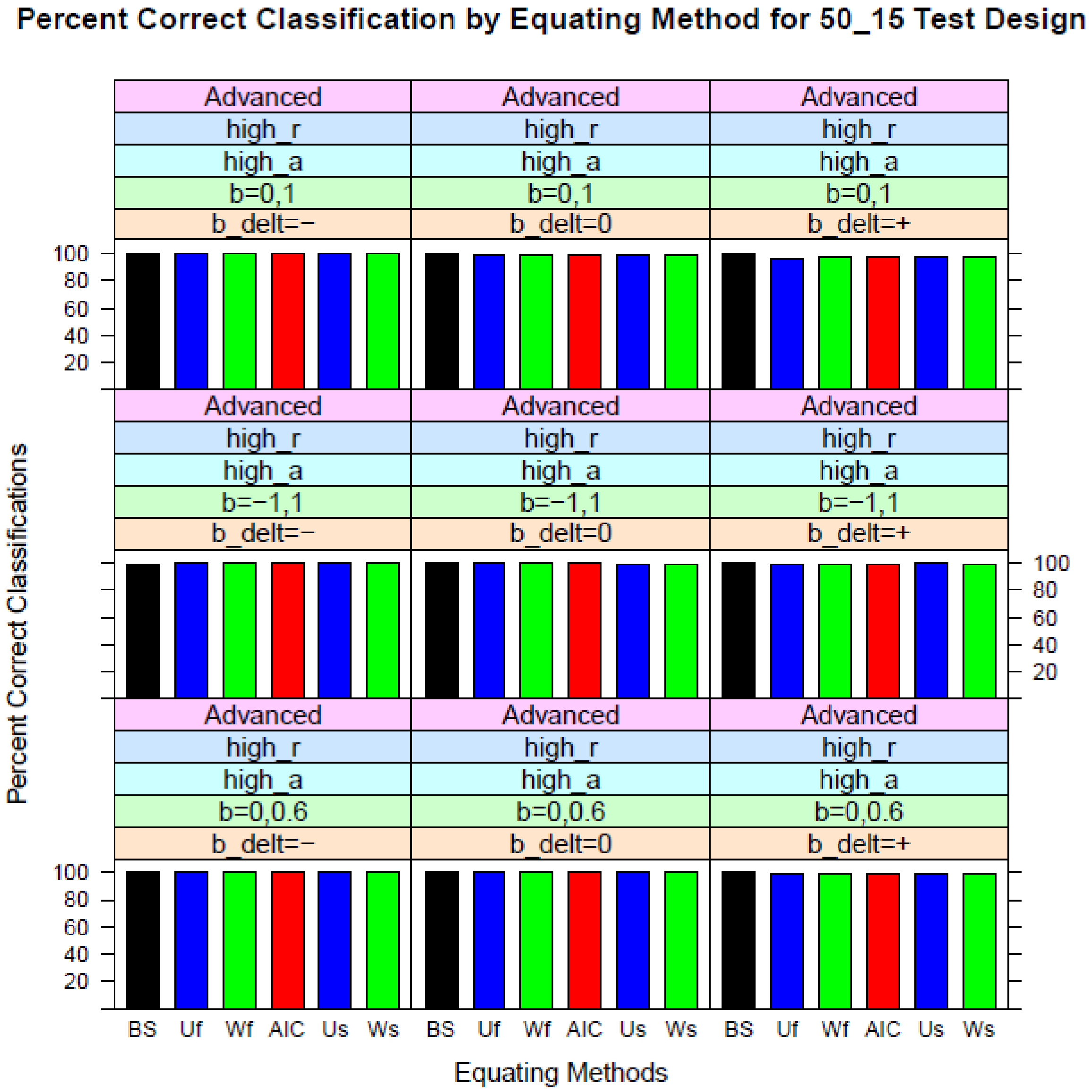


Figure 64. Bar charts showing percent classifications by equating method. (B = Base form, U = unweighted mean equating, W = weighted mean equating, A = Anchor Item Calibration, U_s = stabilized unweighted mean equating with stabilization and W_s = stabilized weighted mean equating with stabilization)

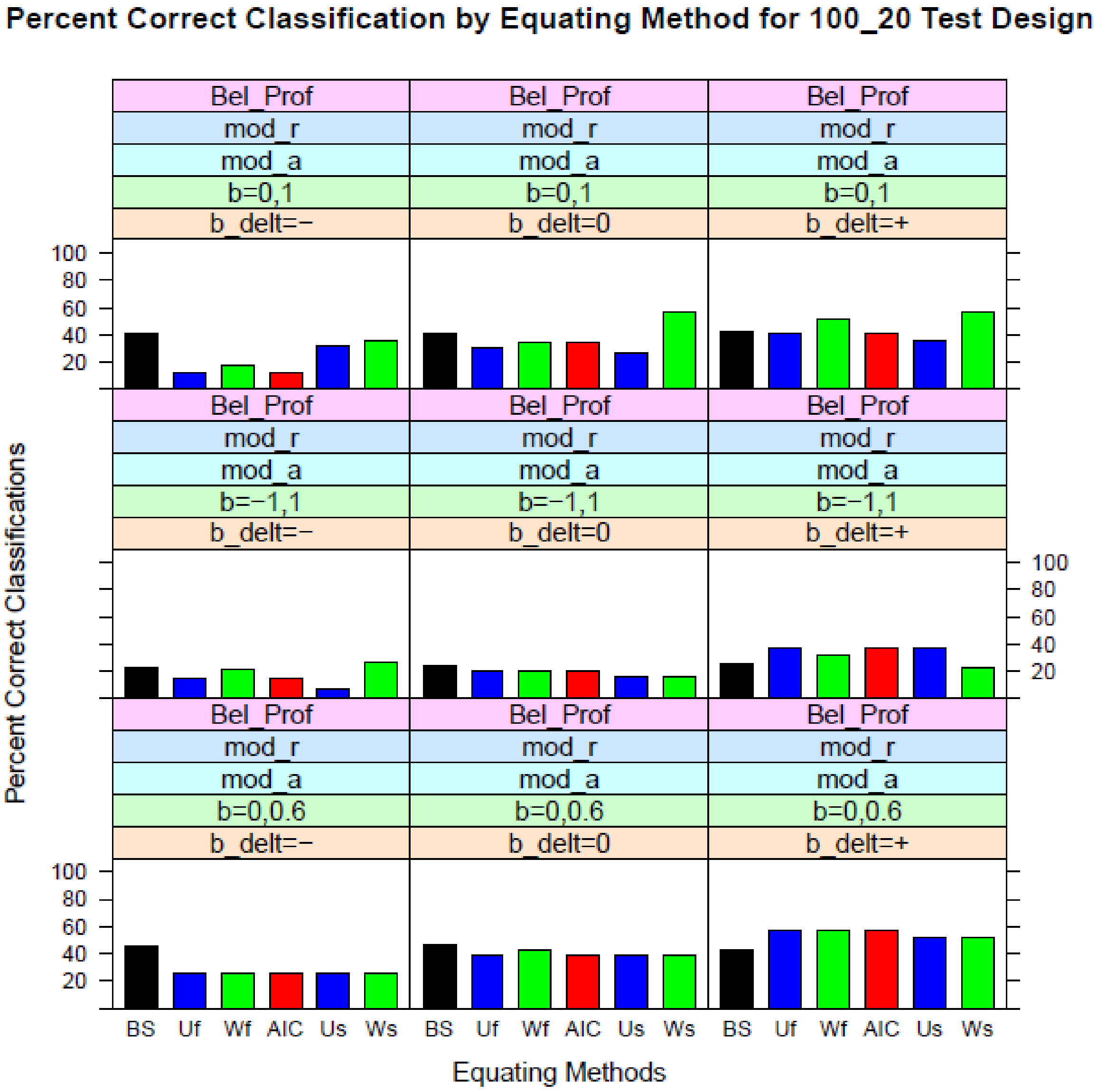


Figure 65. Bar charts showing percent classifications by equating method. (B = Base form, U = unweighted mean equating, W = weighted mean equating, A = Anchor Item Calibration, U_s = stabilized unweighted mean equating with stabilization and W_s = stabilized weighted mean equating with stabilization)

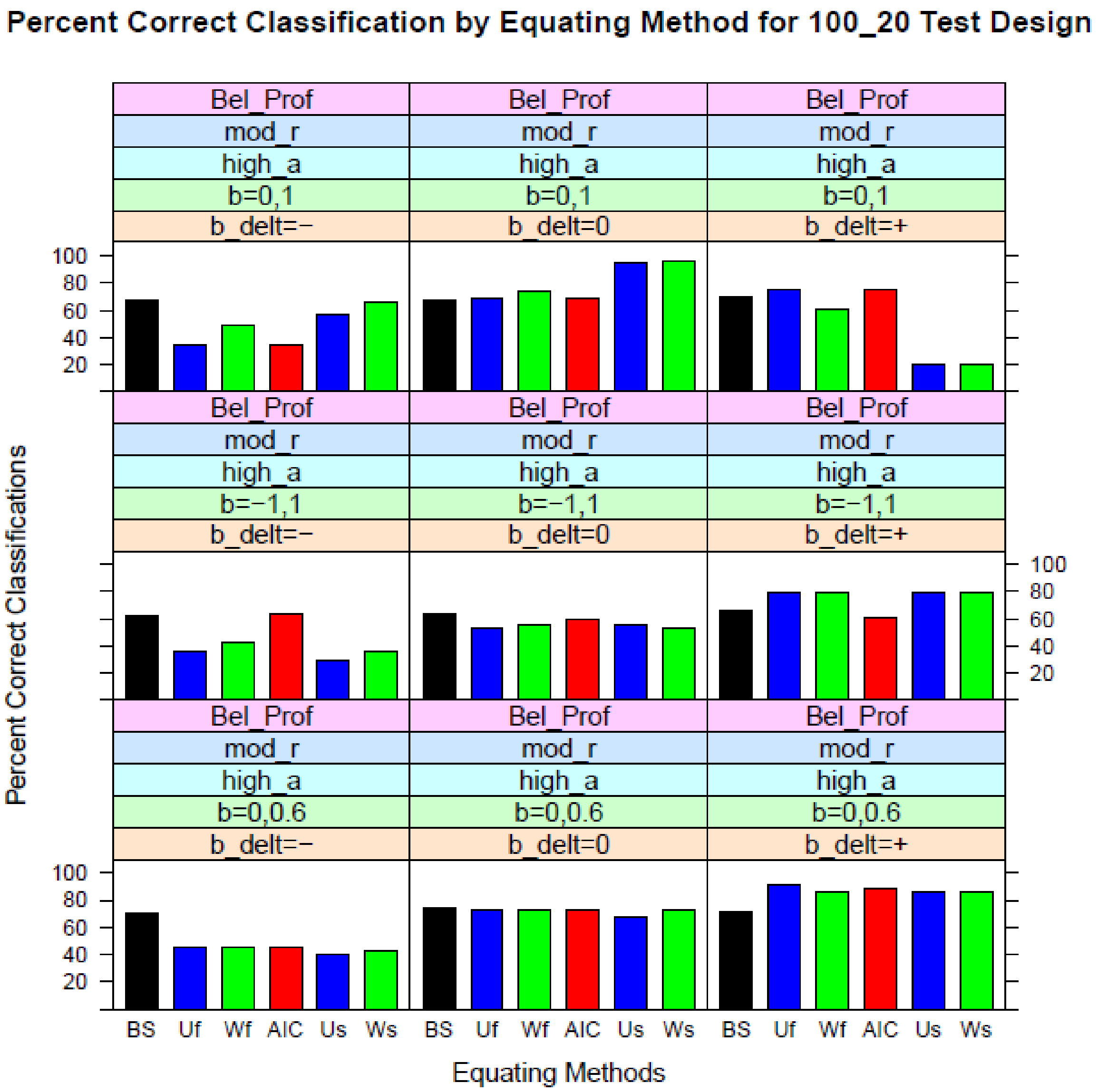


Figure 66. Bar charts showing percent classifications by equating method. (B = Base form, U = unweighted mean equating, W = weighted mean equating, A = Anchor Item Calibration, U_s = stabilized unweighted mean equating with stabilization and W_s = stabilized weighted mean equating with stabilization)

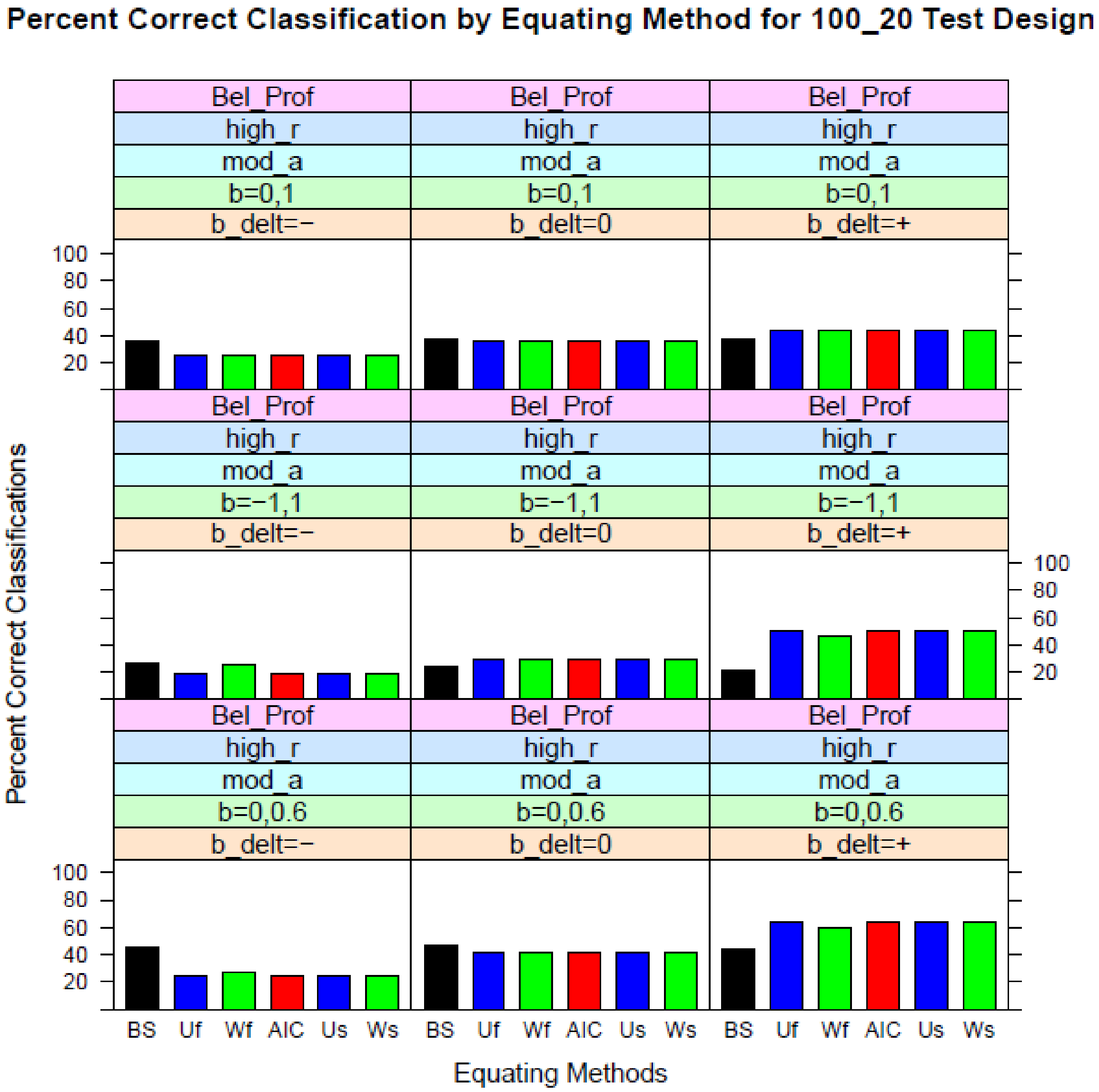


Figure 67. Bar charts showing percent classifications by equating method. (B = Base form, U = unweighted mean equating, W = weighted mean equating, A = Anchor Item Calibration, U_s = stabilized unweighted mean equating with stabilization and W_s = stabilized weighted mean equating with stabilization)

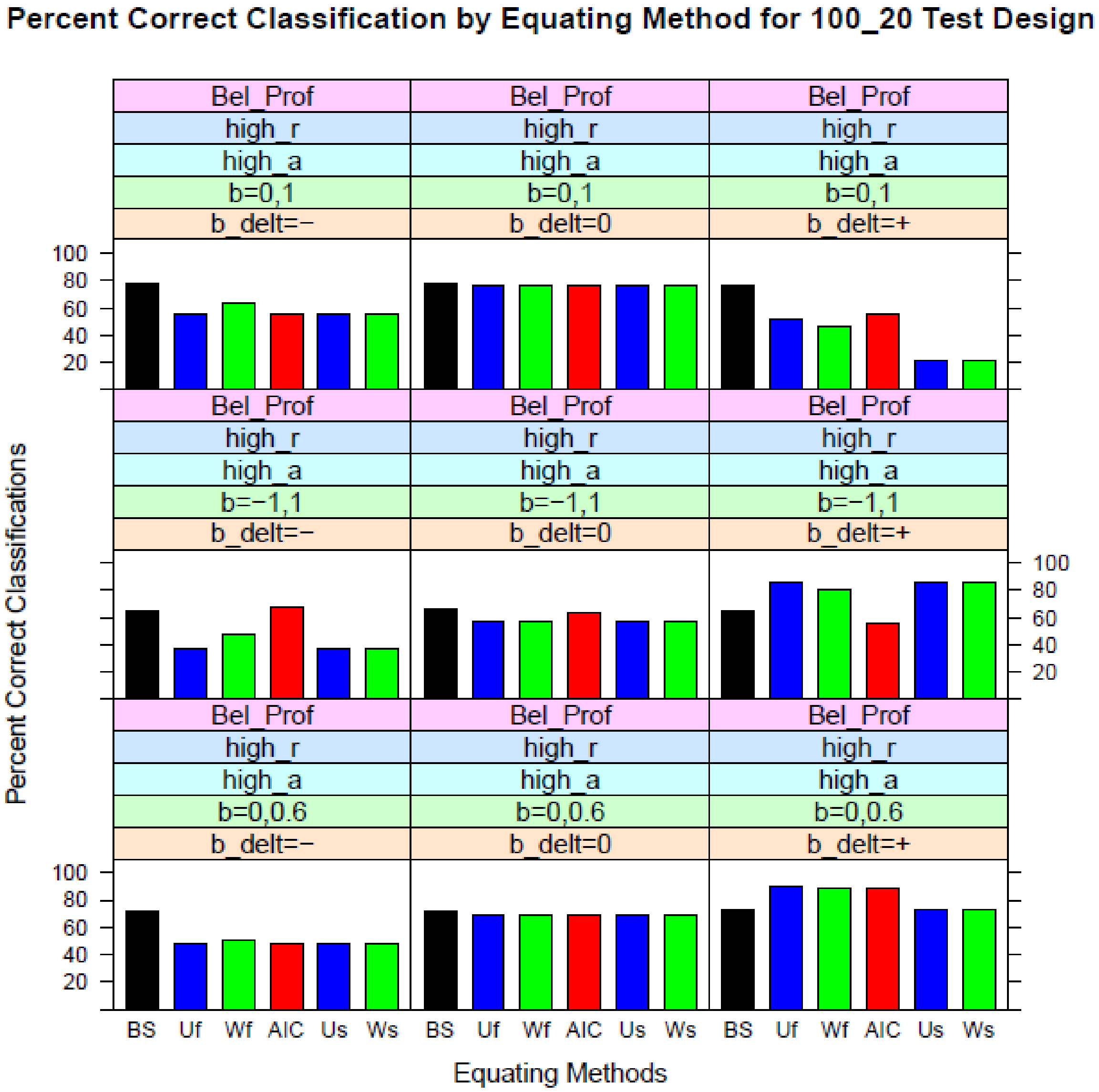


Figure 68. Bar charts showing percent classifications by equating method. (B = Base form, U = unweighted mean equating, W = weighted mean equating, A = Anchor Item Calibration, U_s = stabilized unweighted mean equating with stabilization and W_s = stabilized weighted mean equating with stabilization)

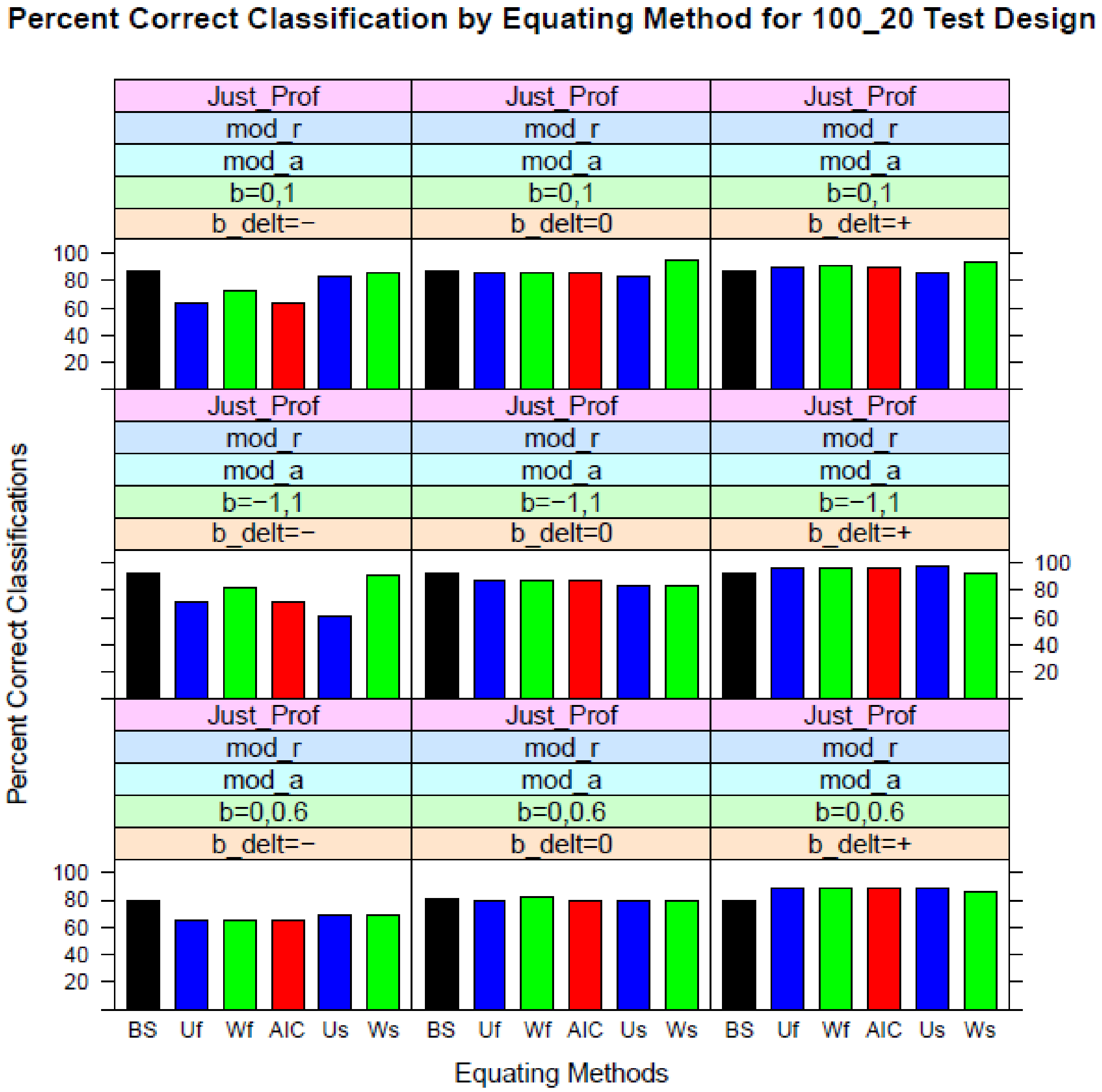


Figure 69. Bar charts showing percent classifications by equating method. (B = Base form, U = unweighted mean equating, W = weighted mean equating, A = Anchor Item Calibration, U_s = stabilized unweighted mean equating with stabilization and W_s = stabilized weighted mean equating with stabilization)

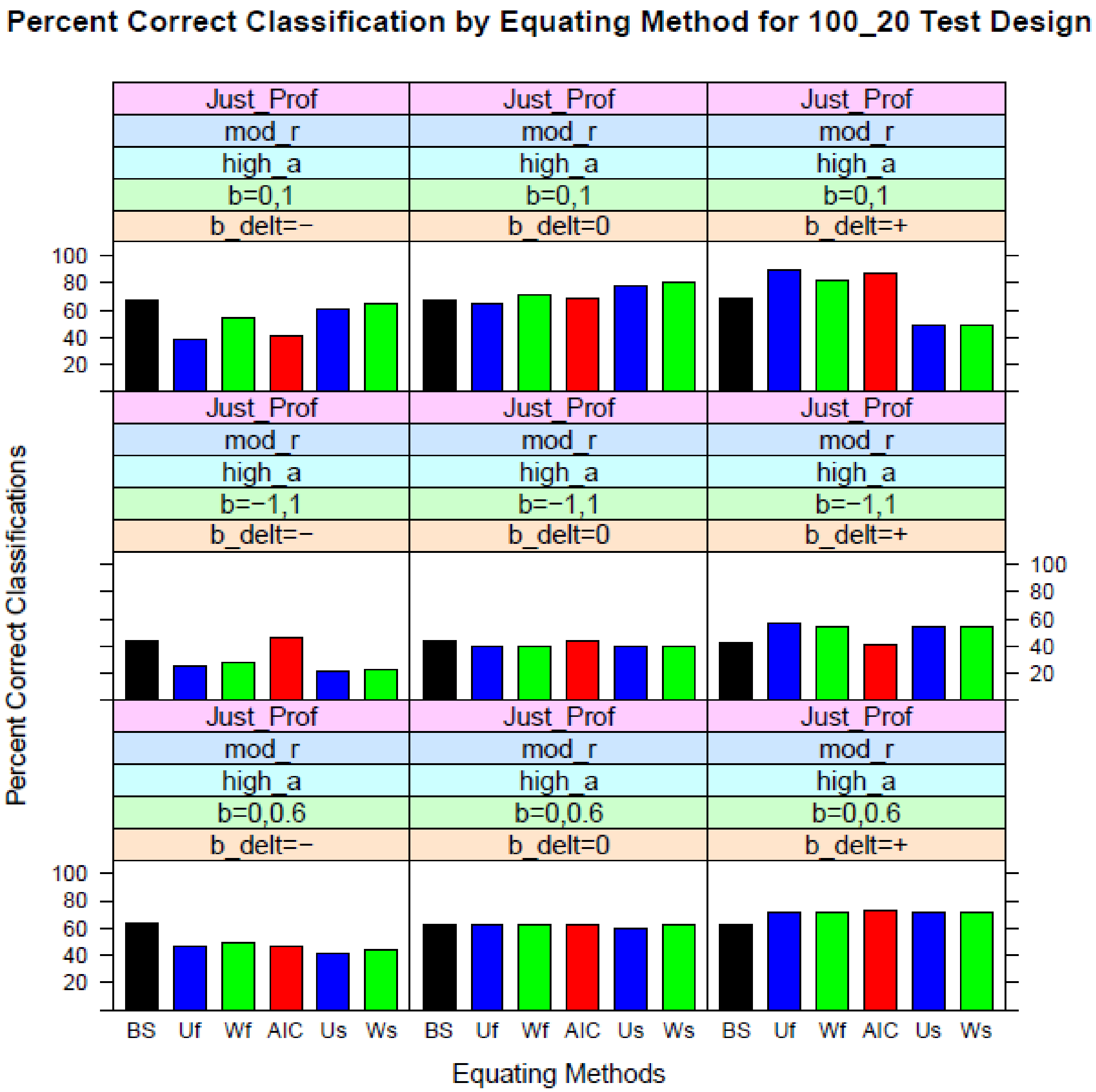


Figure 70. Bar charts showing percent classifications by equating method. (B = Base form, U = unweighted mean equating, W = weighted mean equating, A = Anchor Item Calibration, U_s = stabilized unweighted mean equating with stabilization and W_s = stabilized weighted mean equating with stabilization)

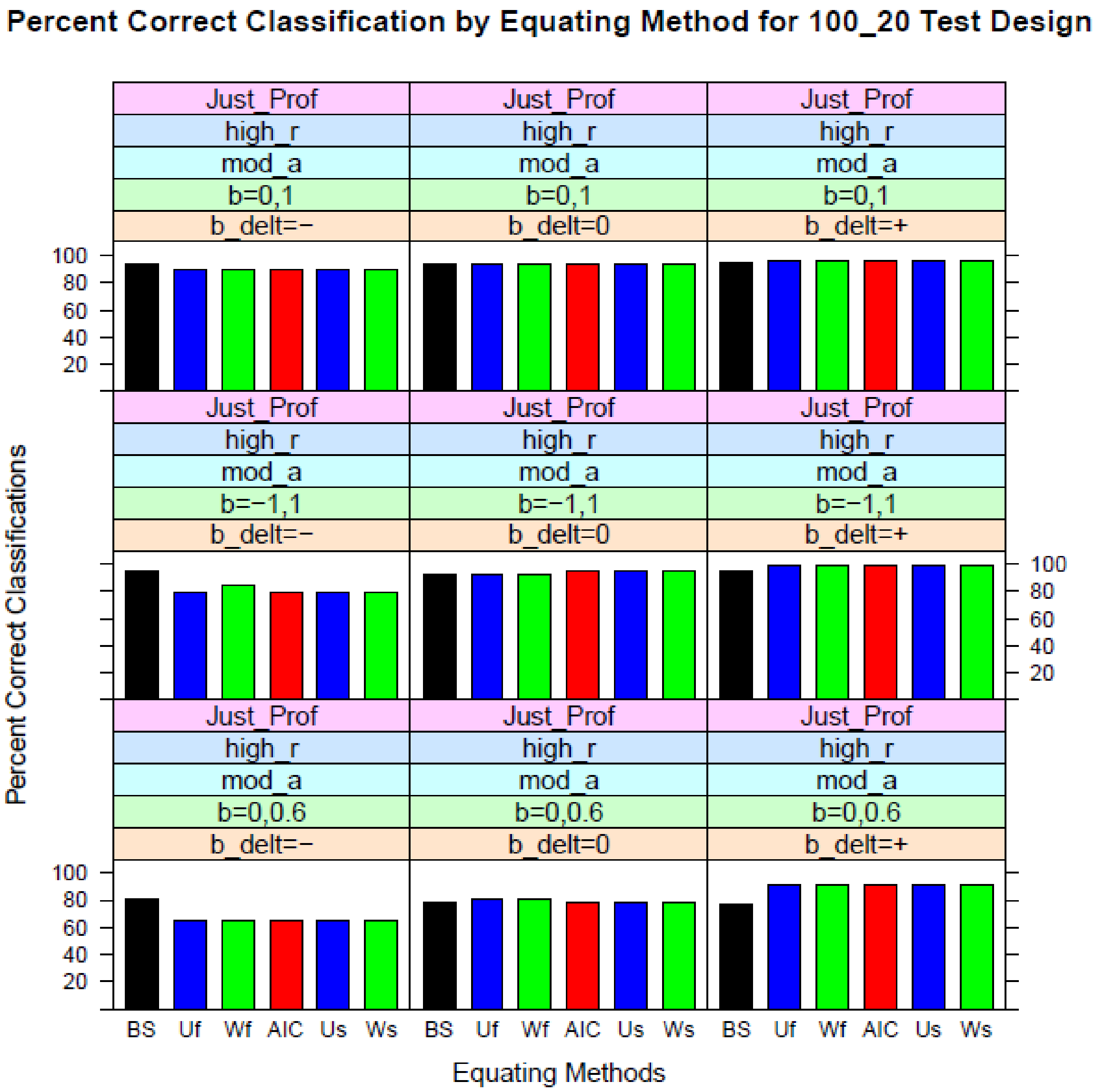


Figure 71. Bar charts showing percent classifications by equating method. (B = Base form, U = unweighted mean equating, W = weighted mean equating, A = Anchor Item Calibration, U_s = stabilized unweighted mean equating with stabilization and W_s = stabilized weighted mean equating with stabilization)

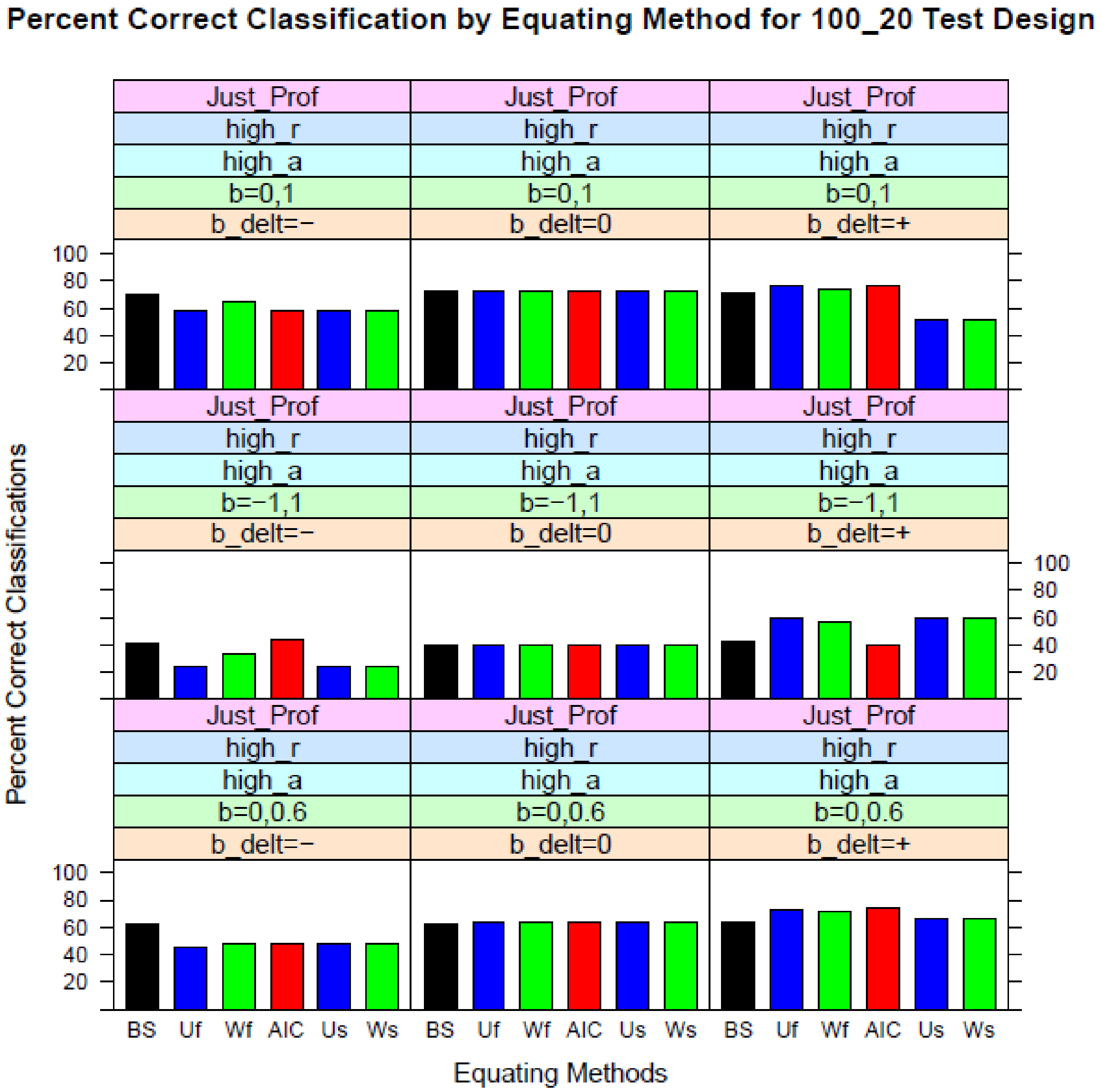


Figure 72. Bar charts showing percent classifications by equating method. (B = Base form, U = unweighted mean equating, W = weighted mean equating, A = Anchor Item Calibration, U_s = stabilized unweighted mean equating with stabilization and W_s = stabilized weighted mean equating with stabilization)

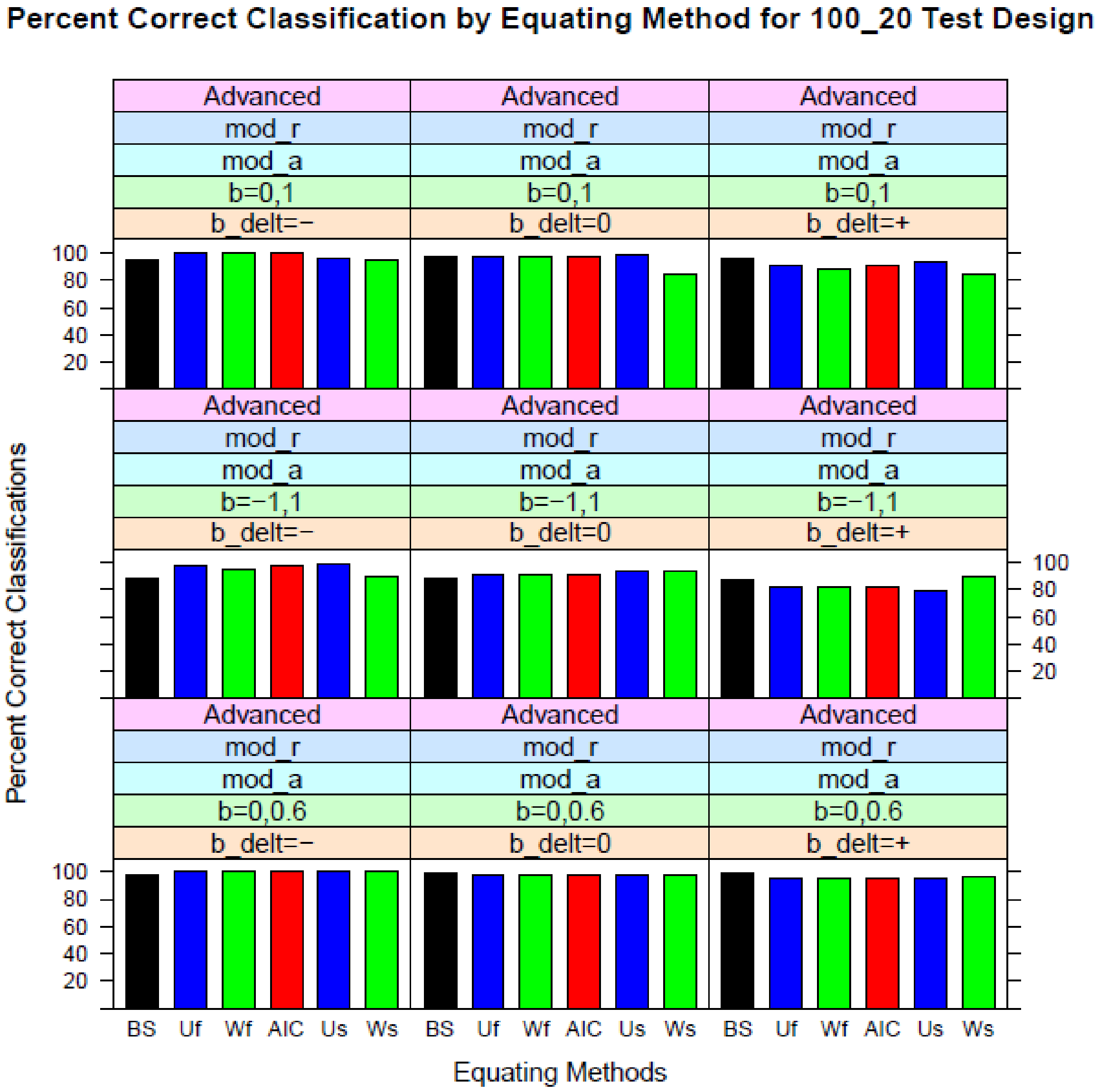


Figure 73. Bar charts showing percent classifications by equating method. (B = Base form, U = unweighted mean equating, W = weighted mean equating, A = Anchor Item Calibration, U_s = stabilized unweighted mean equating with stabilization and W_s = stabilized weighted mean equating with stabilization)

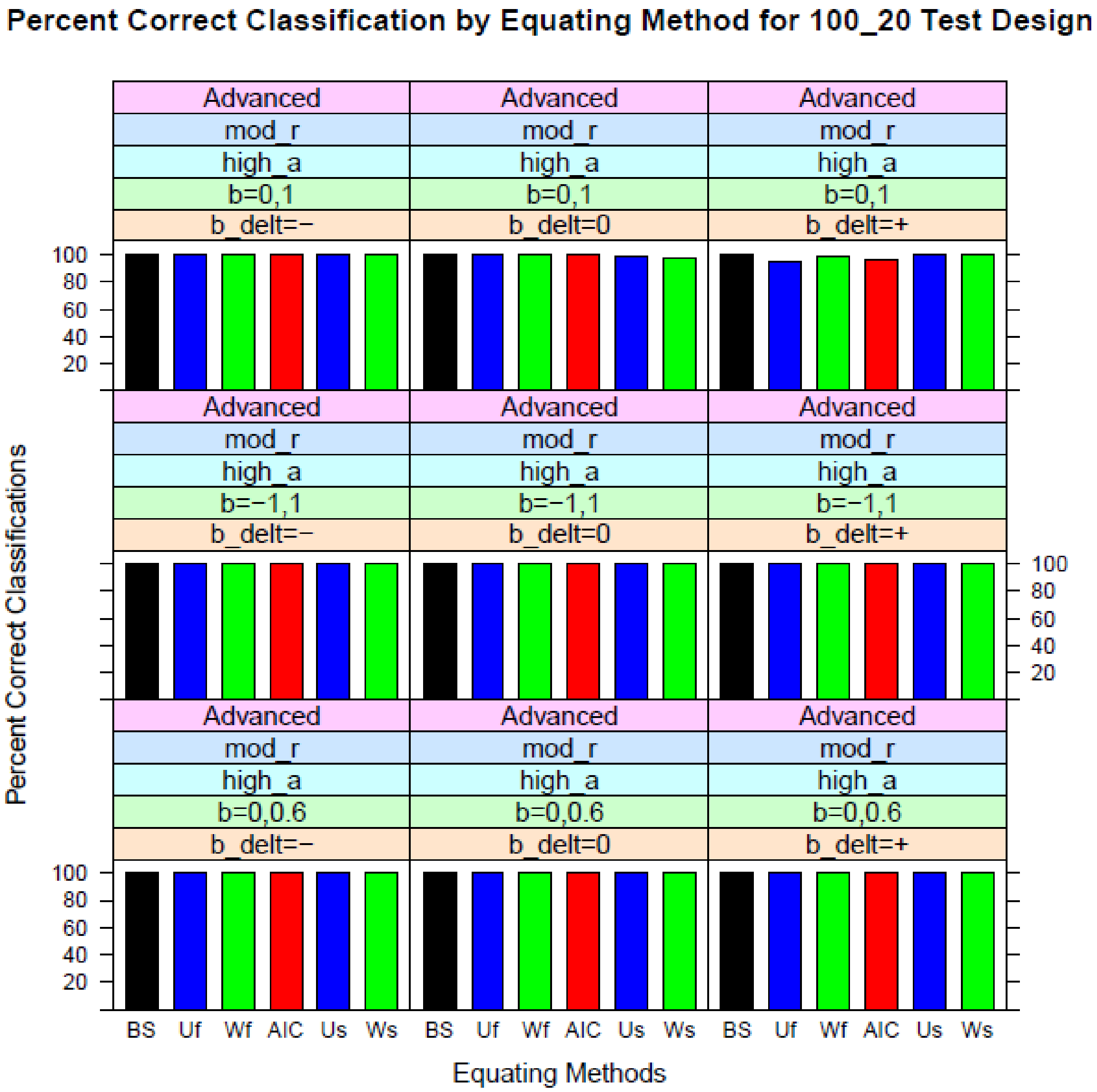


Figure 74. Bar charts showing percent classifications by equating method. (B = Base form, U = unweighted mean equating, W = weighted mean equating, A = Anchor Item Calibration, U_s = stabilized unweighted mean equating with stabilization and W_s = stabilized weighted mean equating with stabilization)

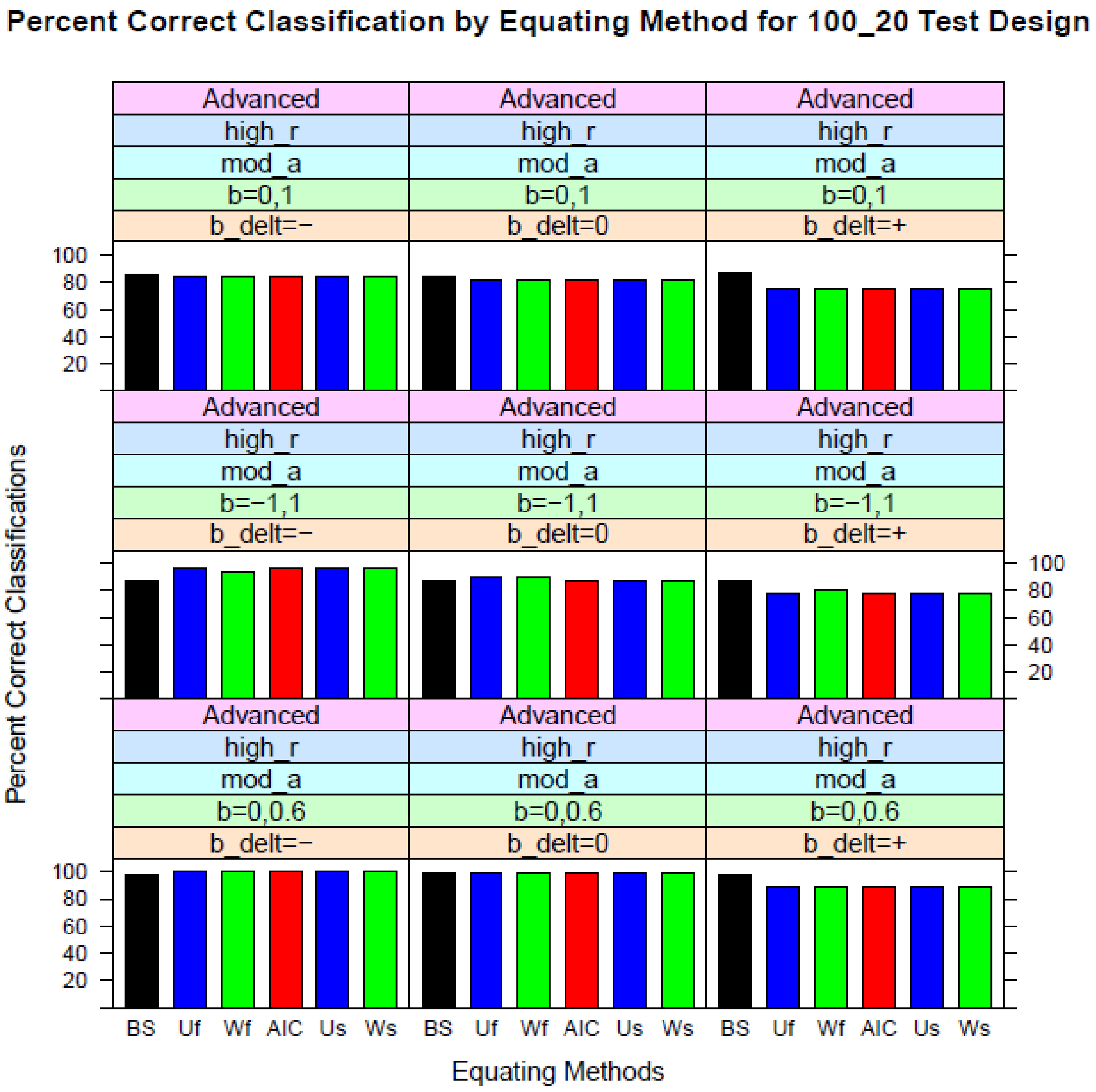


Figure 75. Bar charts showing percent classifications by equating method. (B = Base form, U = unweighted mean equating, W = weighted mean equating, A = Anchor Item Calibration, U_s = stabilized unweighted mean equating with stabilization and W_s = stabilized weighted mean equating with stabilization)

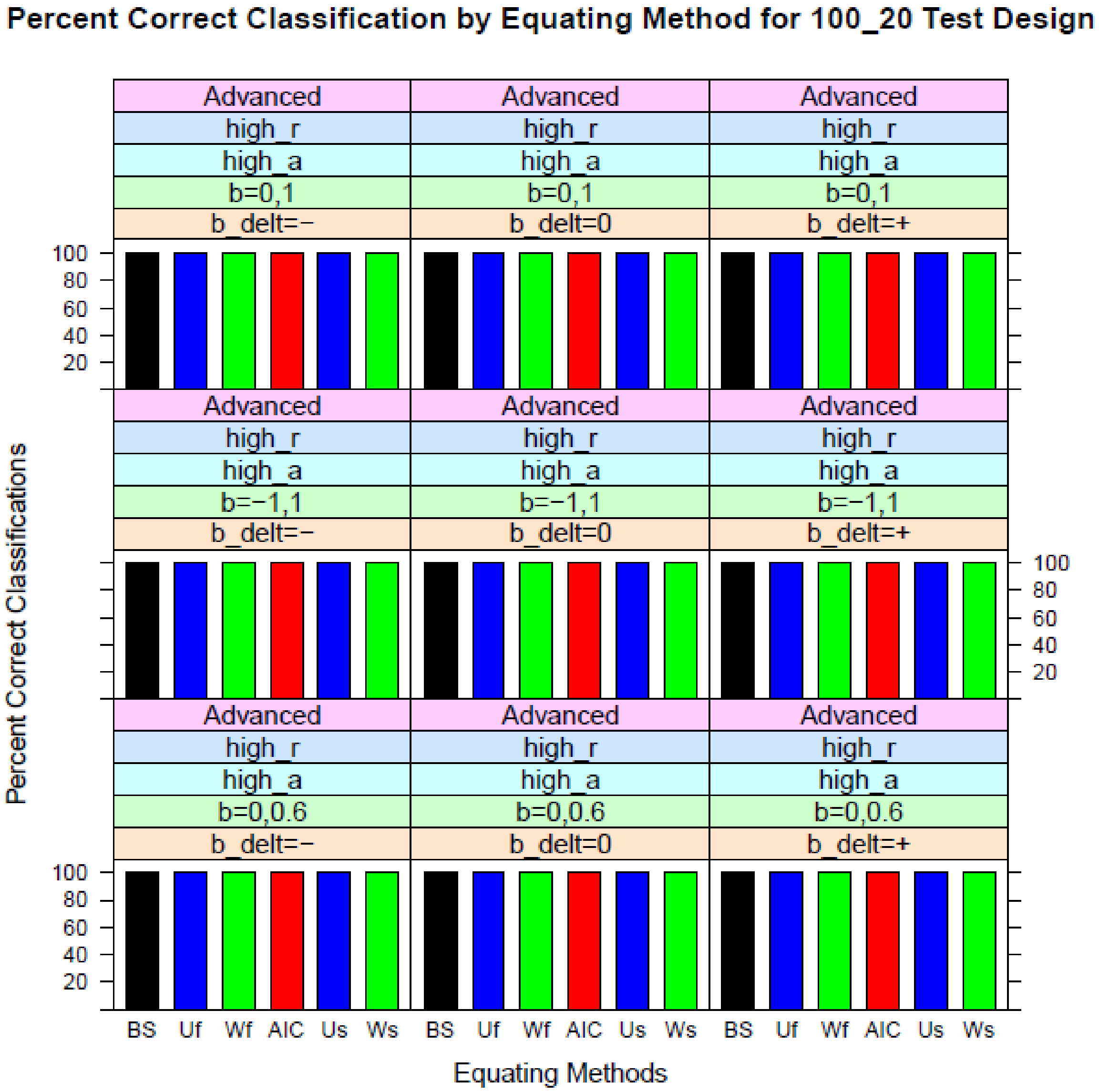
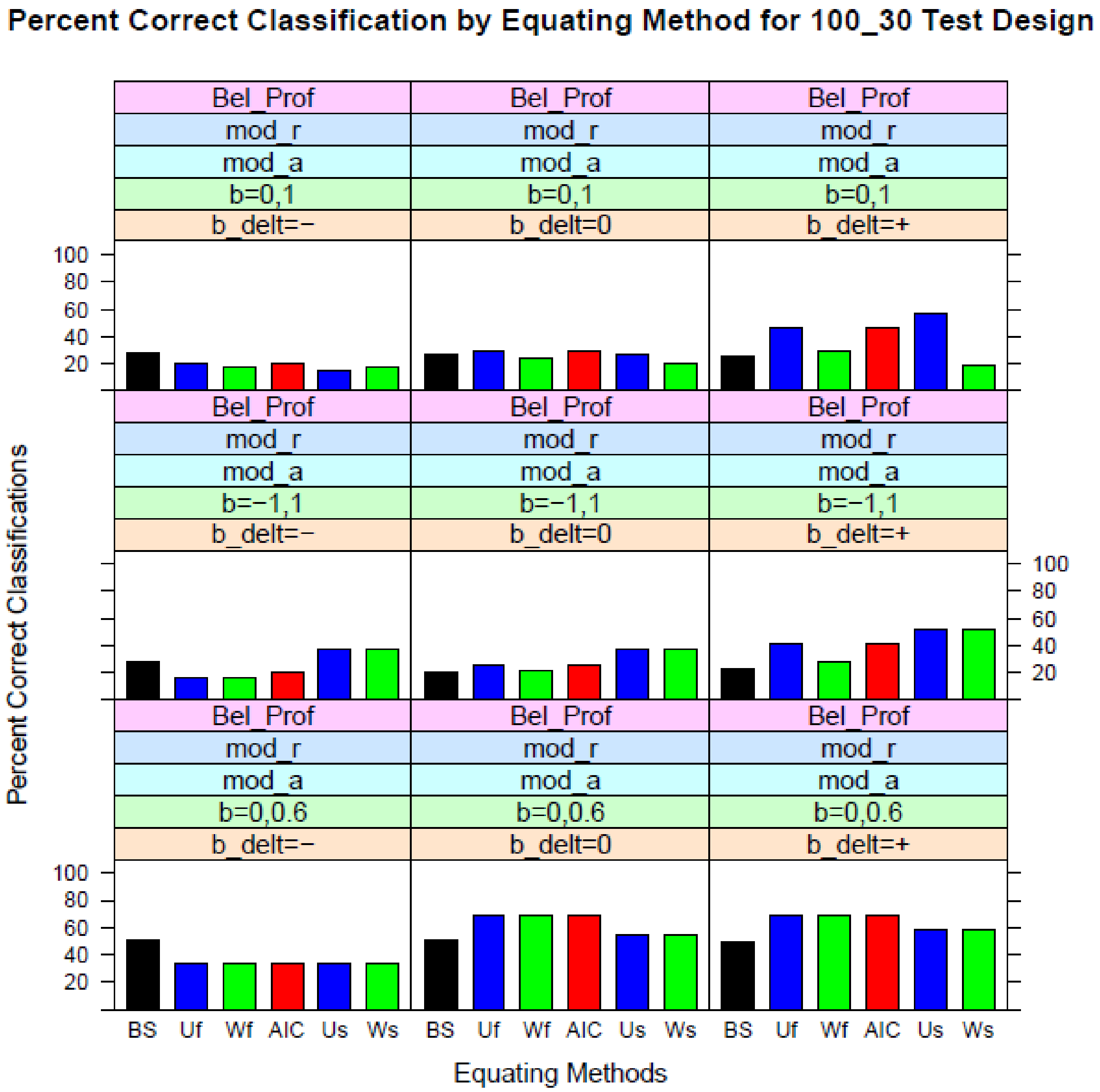


Figure 76. Bar charts showing percent classifications by equating method. (B = Base form, U = unweighted mean equating, W = weighted mean equating, A = Anchor Item Calibration, U_s = stabilized unweighted mean equating with stabilization and W_s = stabilized weighted mean equating with stabilization)



Percent Correct Classification by Equating Method for 100_30 Test Design

Figure 77. Bar charts showing percent classifications by equating method. (B = Base form, U = unweighted mean equating, W = weighted mean equating, A = Anchor Item Calibration, U_s = stabilized unweighted mean equating with stabilization and W_s = stabilized weighted mean equating with stabilization)

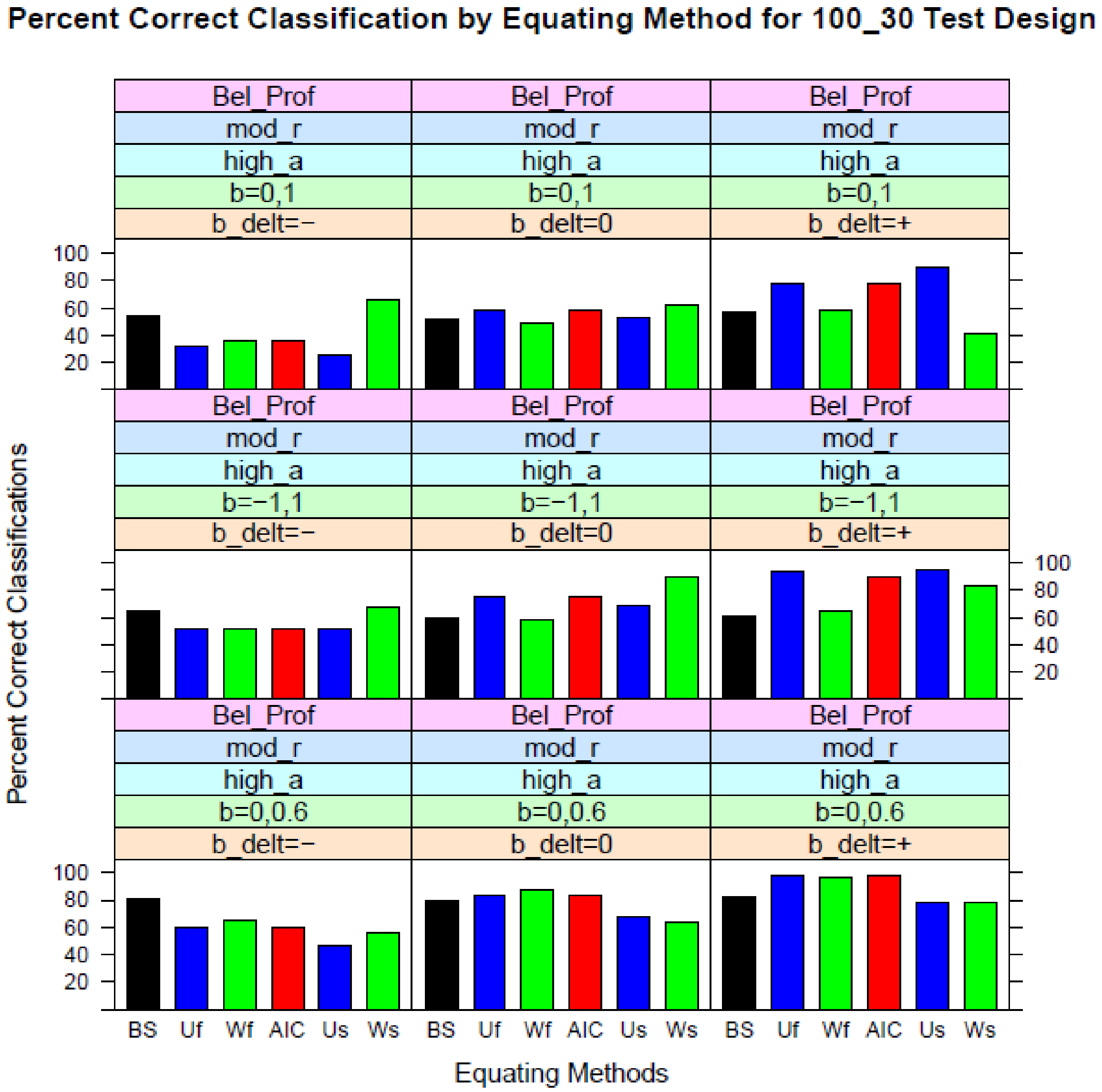


Figure 78. Bar charts showing percent classifications by equating method. (B = Base form, U = unweighted mean equating, W = weighted mean equating, A = Anchor Item Calibration, U_s = stabilized unweighted mean equating with stabilization and W_s = stabilized weighted mean equating with stabilization)

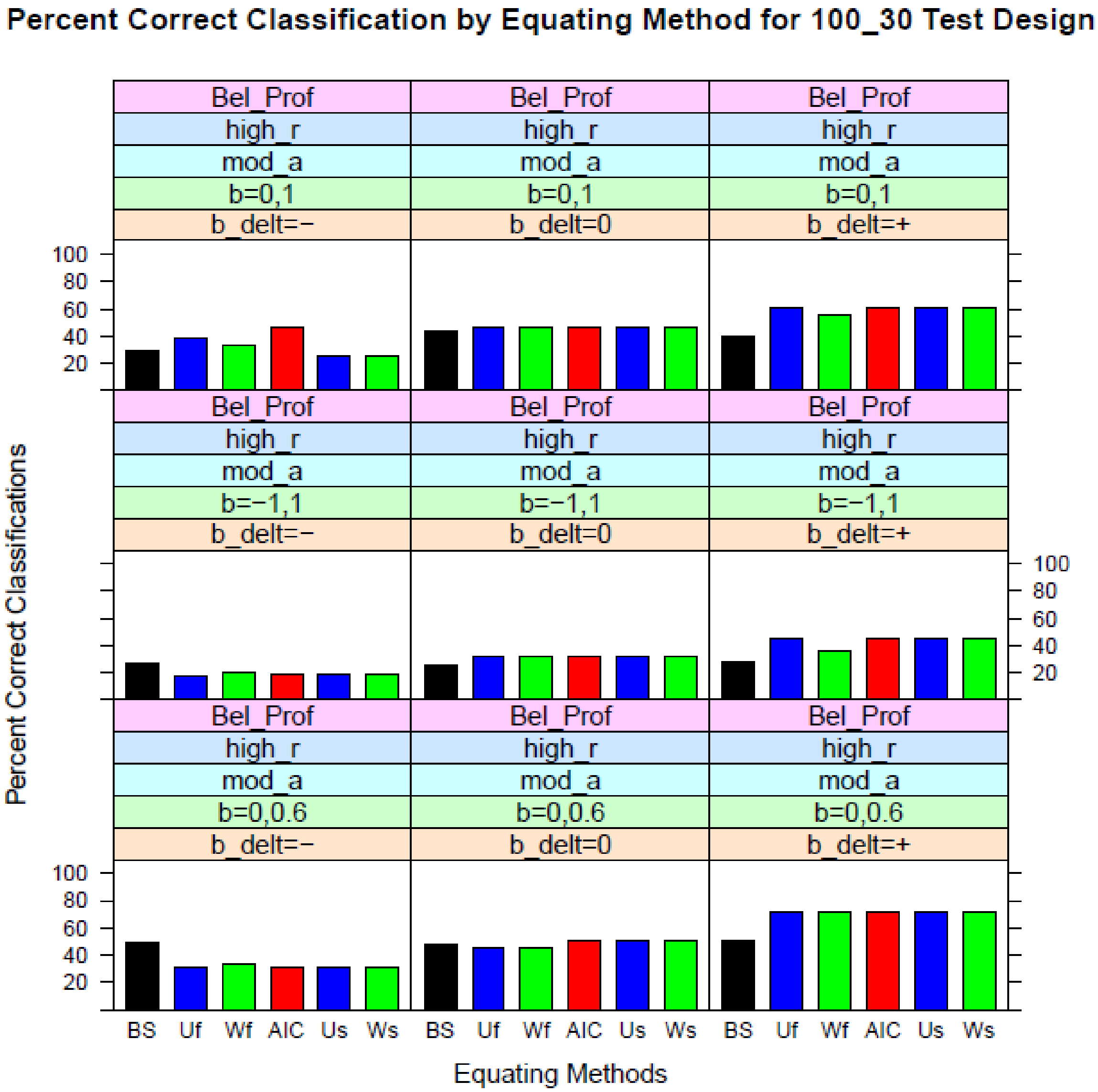


Figure 79. Bar charts showing percent classifications by equating method. (B = Base form, U = unweighted mean equating, W = weighted mean equating, A = Anchor Item Calibration, U_s = stabilized unweighted mean equating with stabilization and W_s = stabilized weighted mean equating with stabilization)

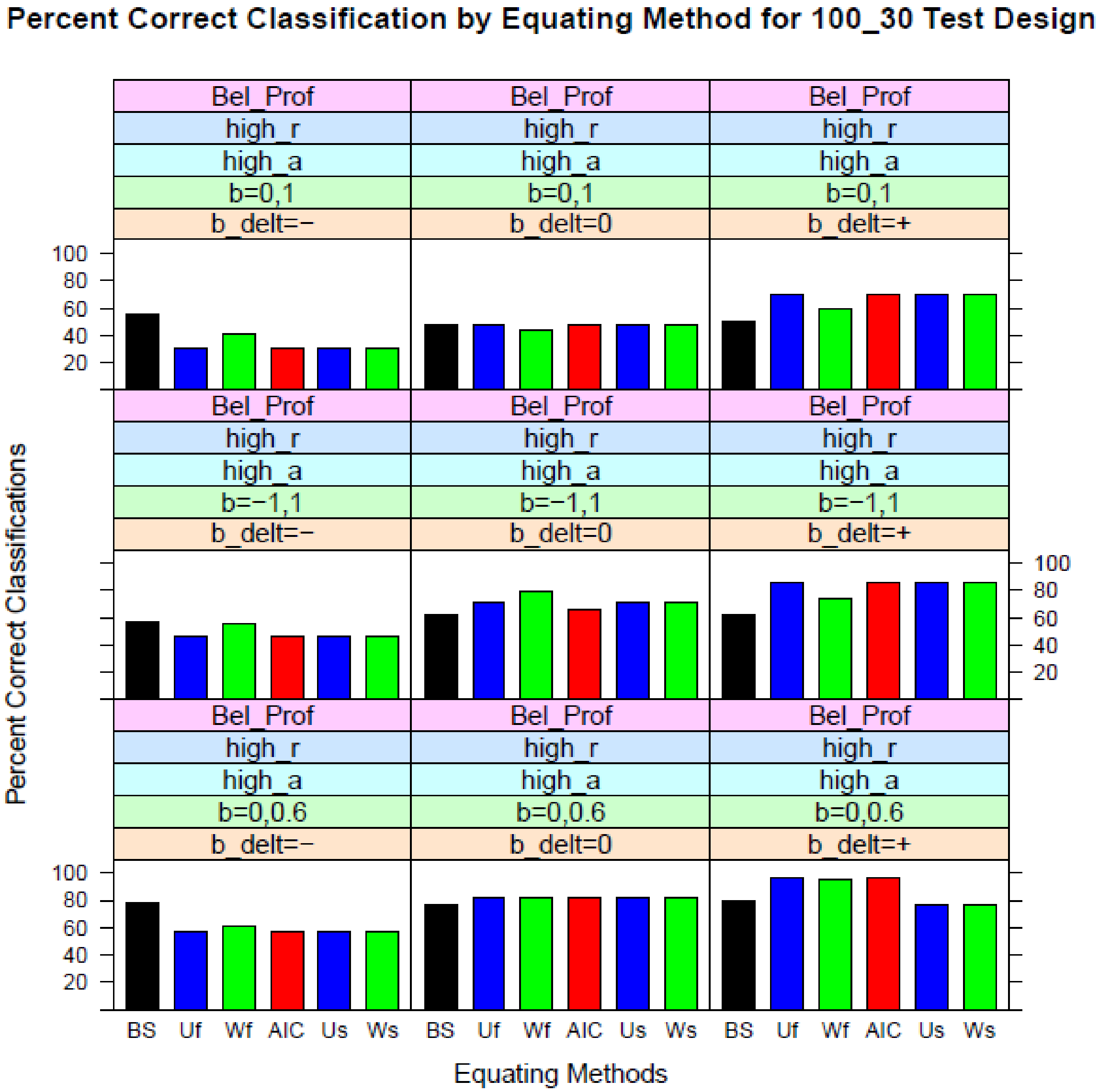


Figure 80. Bar charts showing percent classifications by equating method. (B = Base form, U = unweighted mean equating, W = weighted mean equating, A = Anchor Item Calibration, U_s = stabilized unweighted mean equating with stabilization and W_s = stabilized weighted mean equating with stabilization)

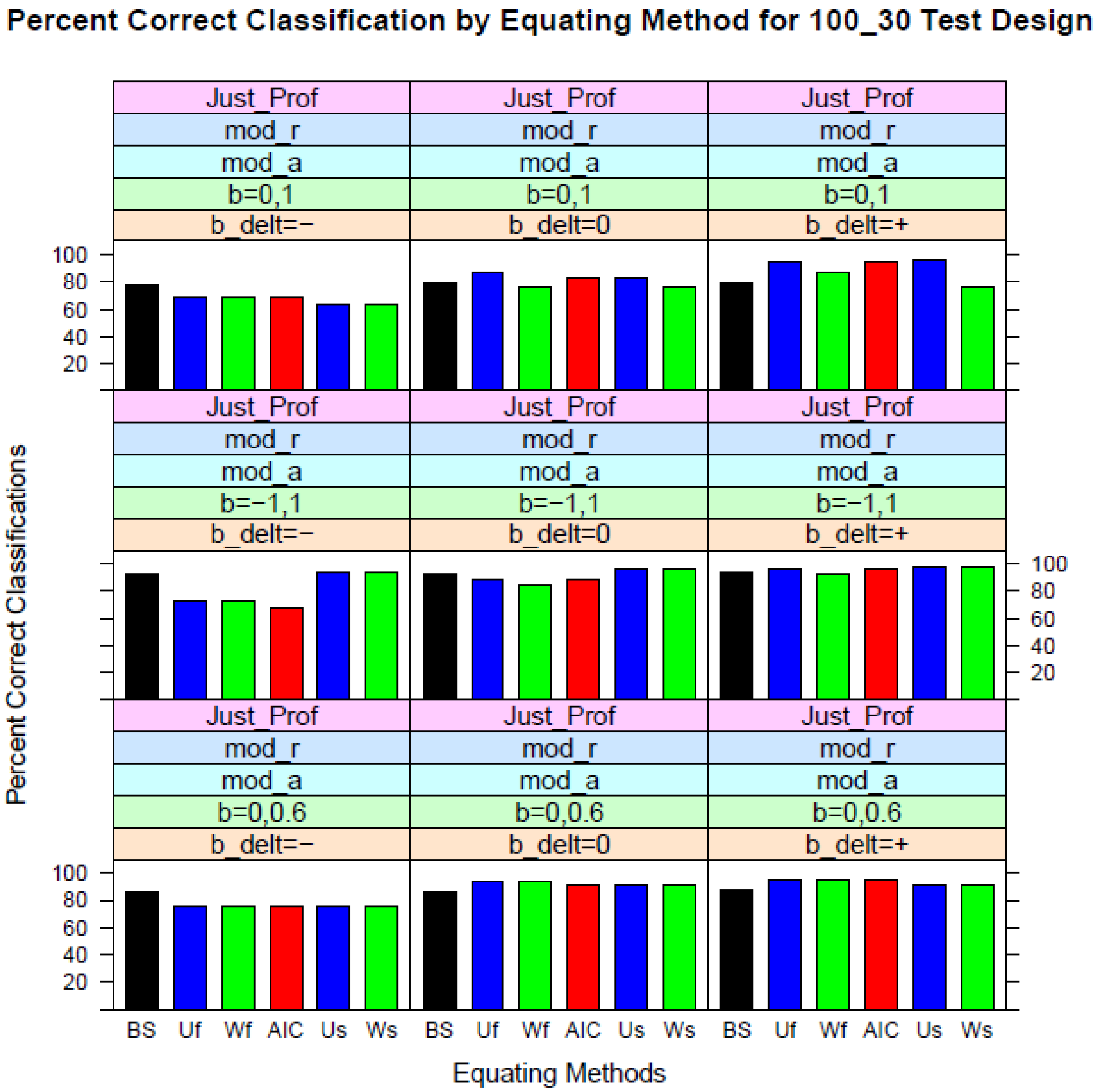


Figure 81. Bar charts showing percent classifications by equating method. (B = Base form, U = unweighted mean equating, W = weighted mean equating, A = Anchor Item Calibration, U_s = stabilized unweighted mean equating with stabilization and W_s = stabilized weighted mean equating with stabilization)

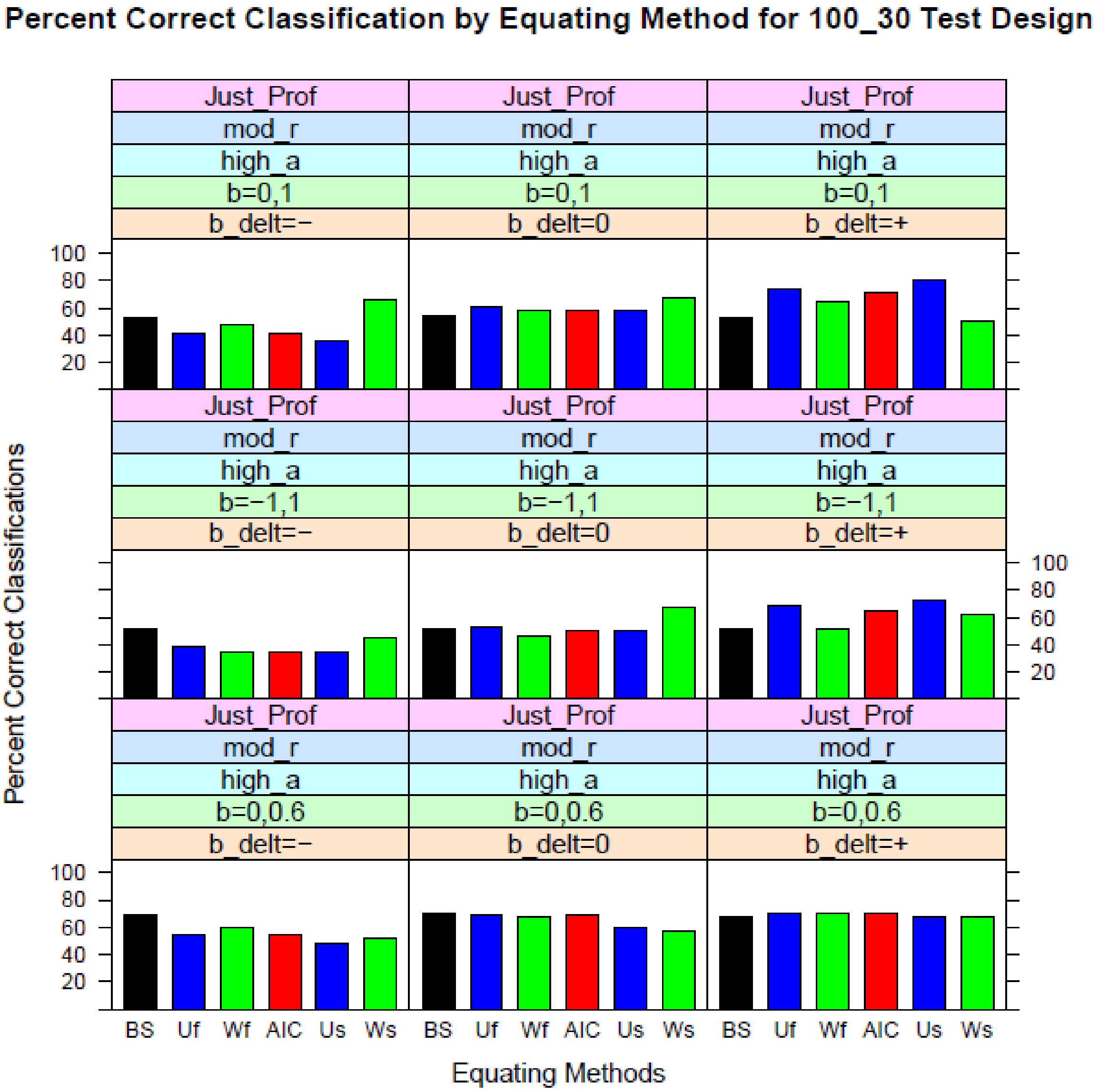


Figure 82. Bar charts showing percent classifications by equating method. (B = Base form, U = unweighted mean equating, W = weighted mean equating, A = Anchor Item Calibration, U_s = stabilized unweighted mean equating with stabilization and W_s = stabilized weighted mean equating with stabilization)

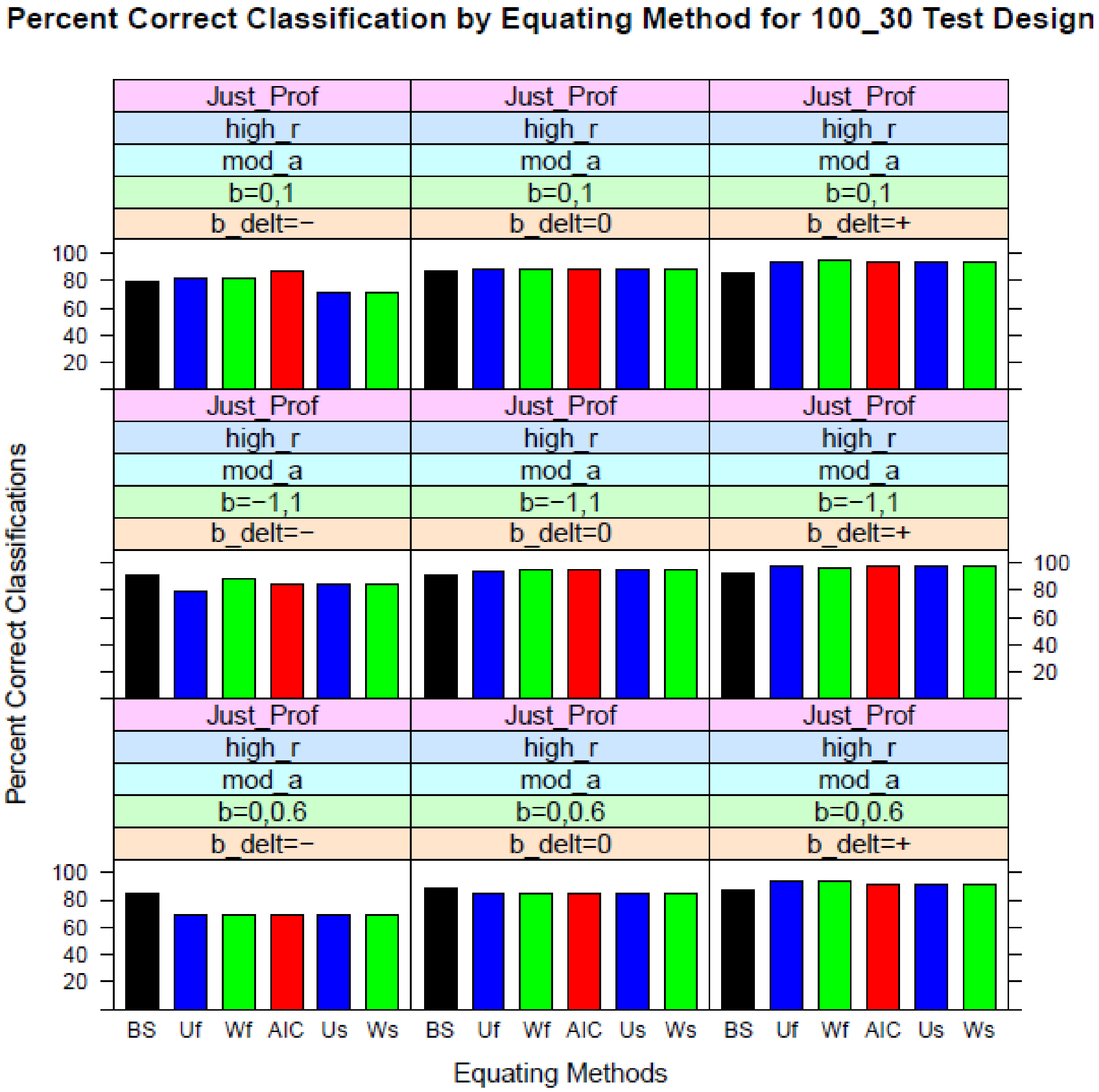


Figure 83. Bar charts showing percent classifications by equating method. (B = Base form, U = unweighted mean equating, W = weighted mean equating, A = Anchor Item Calibration, U_s = stabilized unweighted mean equating with stabilization and W_s = stabilized weighted mean equating with stabilization)

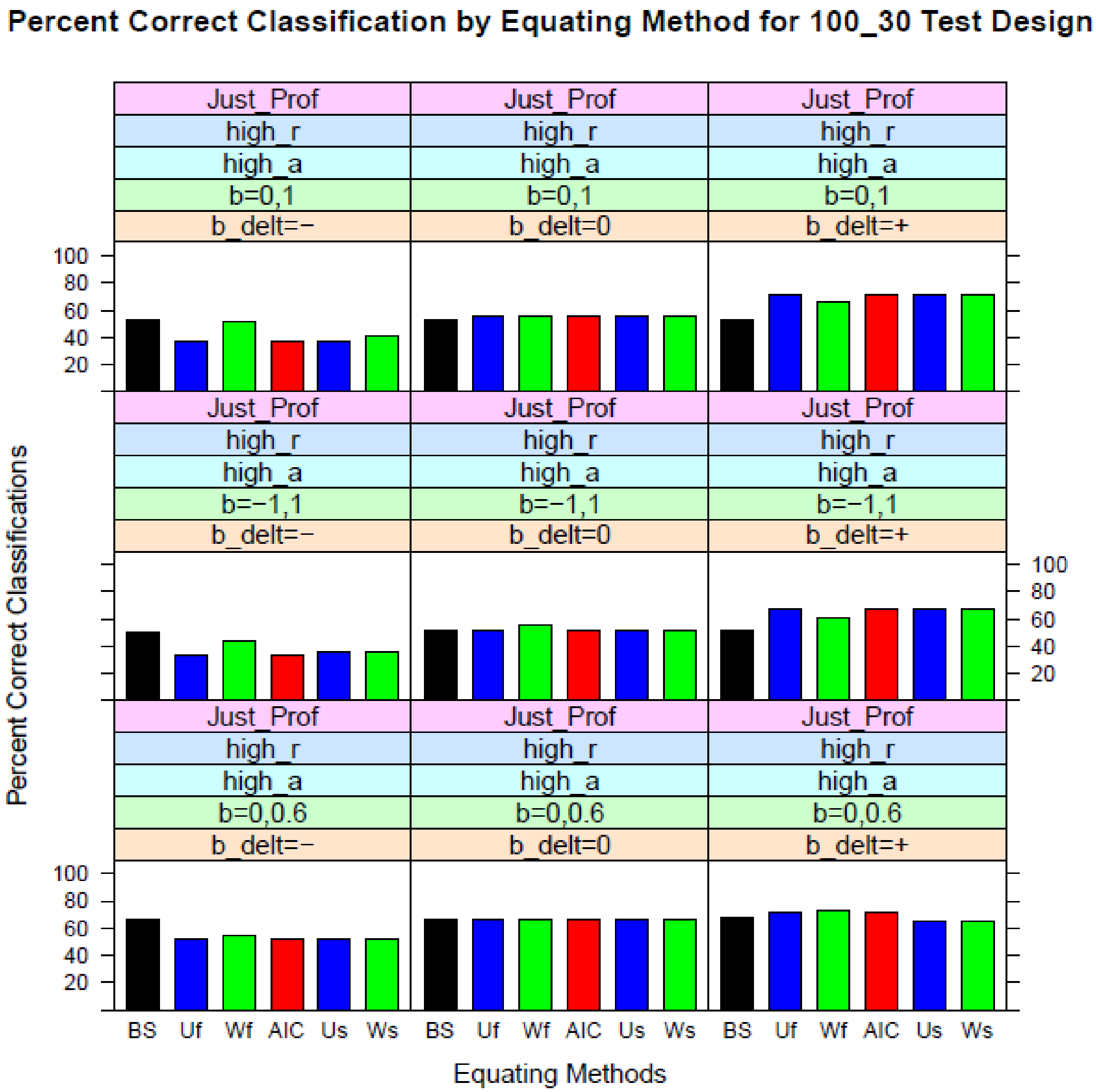


Figure 84. Bar charts showing percent classifications by equating method. (B = Base form, U = unweighted mean equating, W = weighted mean equating, A = Anchor Item Calibration, U_s = stabilized unweighted mean equating with stabilization and W_s = stabilized weighted mean equating with stabilization)

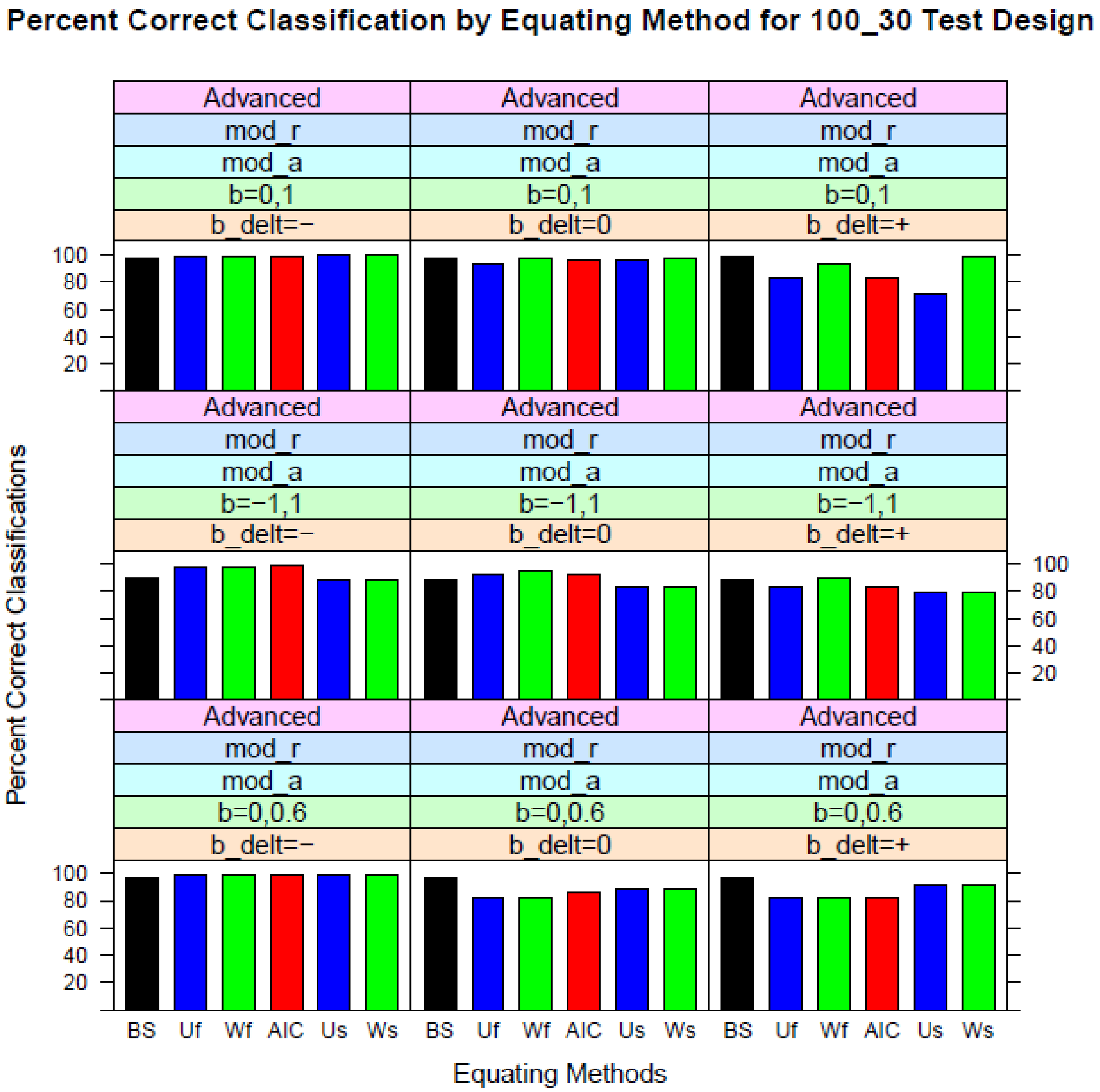


Figure 85. Bar charts showing percent classifications by equating method. (B = Base form, U = unweighted mean equating, W = weighted mean equating, A = Anchor Item Calibration, U_s = stabilized unweighted mean equating with stabilization and W_s = stabilized weighted mean equating with stabilization)

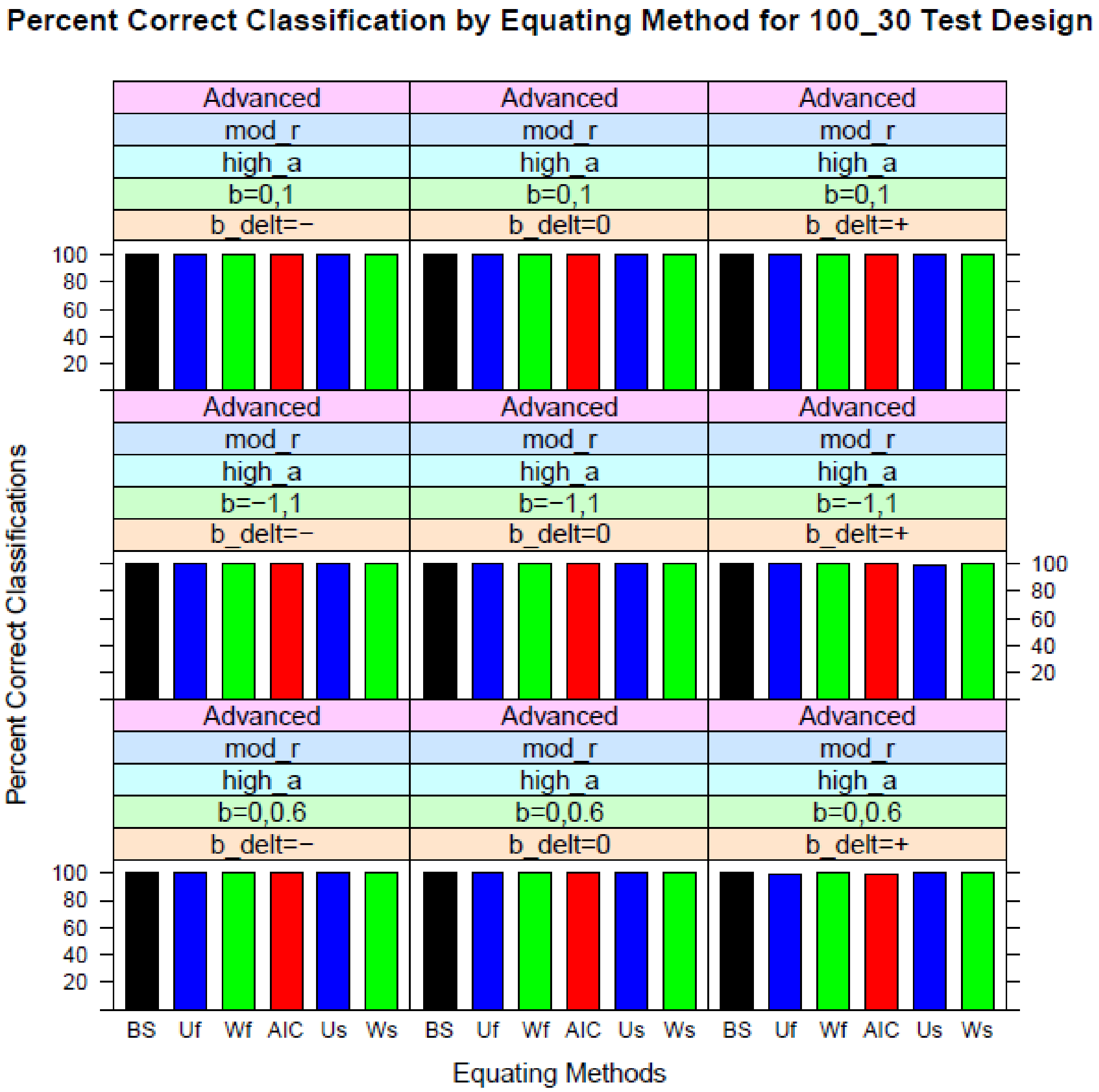


Figure 86. Bar charts showing percent classifications by equating method. (B = Base form, U = unweighted mean equating, W = weighted mean equating, A = Anchor Item Calibration, U_s = stabilized unweighted mean equating with stabilization and W_s = stabilized weighted mean equating with stabilization)

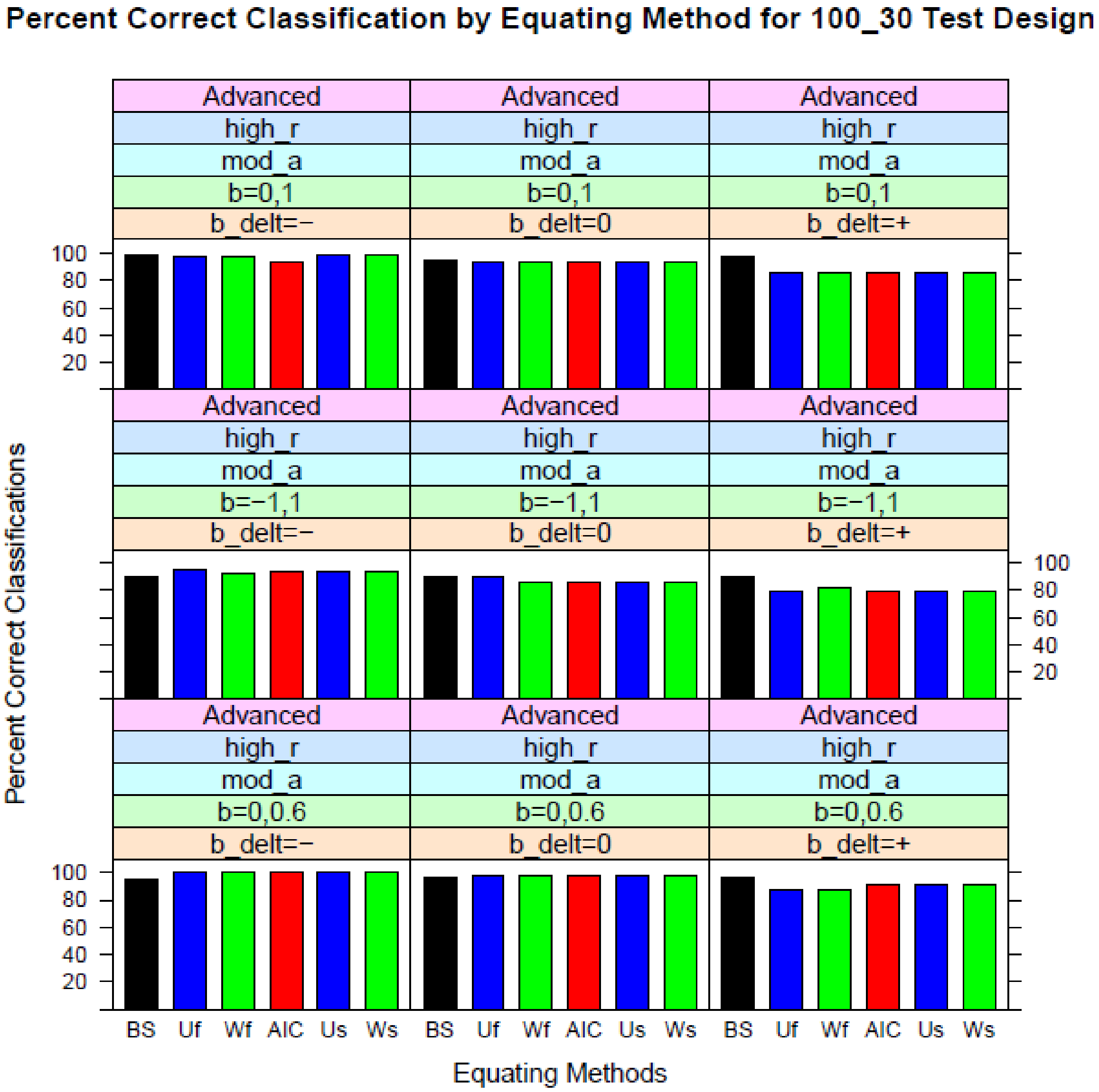


Figure 87. Bar charts showing percent classifications by equating method. (B = Base form, U = unweighted mean equating, W = weighted mean equating, A = Anchor Item Calibration, U_s = stabilized unweighted mean equating with stabilization and W_s = stabilized weighted mean equating with stabilization)

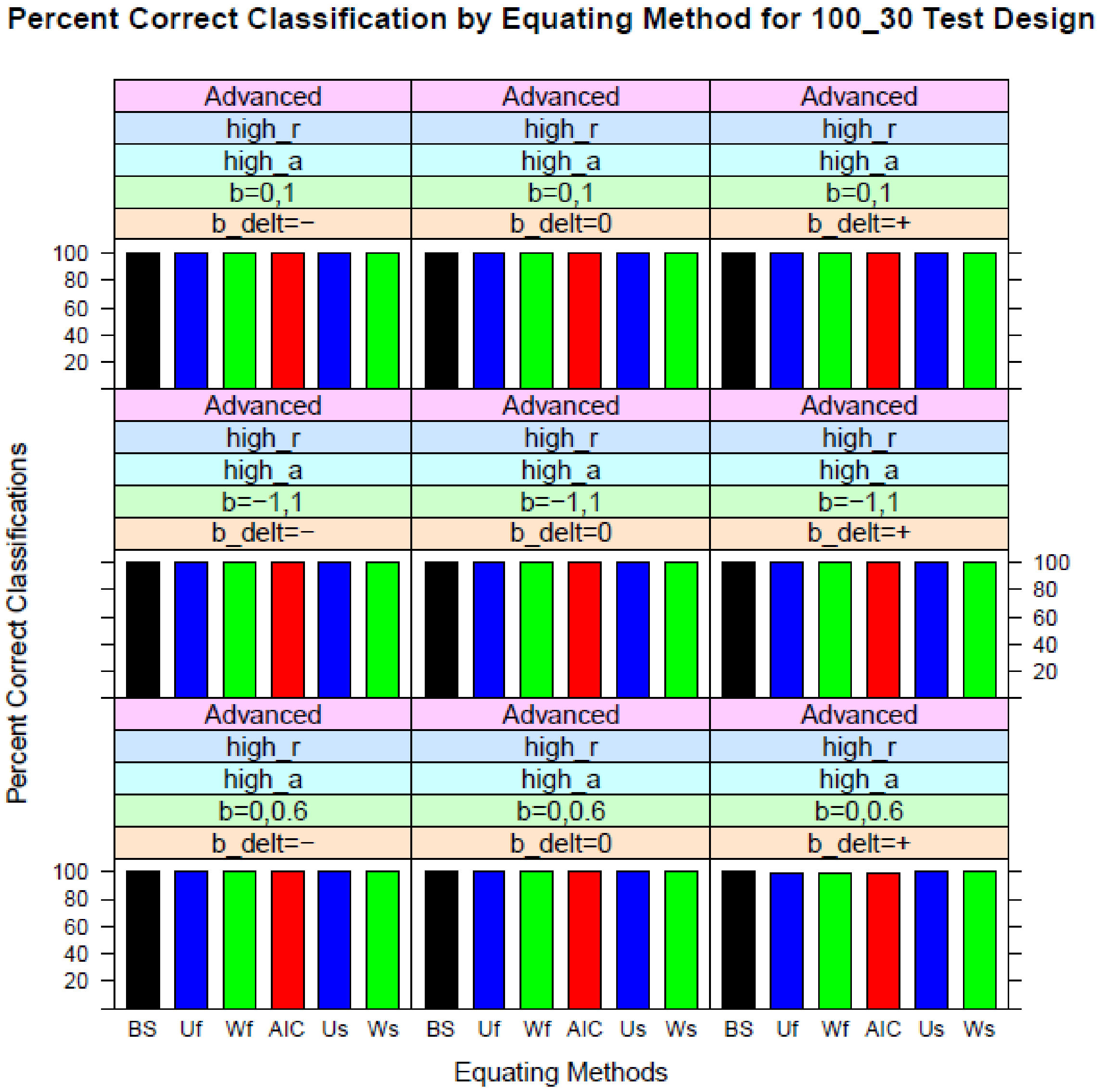


Figure 88. Bar charts showing percent classifications by equating method (B = Base form, U = unweighted mean equating, W = weighted mean equating, A = Anchor Item Calibration, U_s = stabilized unweighted mean equating with stabilization and W_s = stabilized weighted mean equating with stabilization)