An Item Response Theory Approach to Causal Inference in the Presence of a Pre-intervention Assessment

Jessica P. Marini

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2013

Abstract

An Item Response Theory Approach to Causal Inference:

in the Presence of a Pre-intervention Assessment

Jessica Marini

This research develops a form of causal inference based on Item Response Theory (IRT) to combat bias that occurs when existing causal inference methods are used under certain scenarios. When a pre-test is administered, prior to a treatment decision, bias can occur in causal inferences about the decision's effect on the outcome. This new IRT based method uses item-level information, treatment placement, and the outcome to produce estimates of each subject's ability in the chosen domain. Examining a causal inference research question in an IRT model-based framework becomes a model-based way to match subjects on estimates of their true ability. This model-based matching allows inferences to be made about a subject's performance as if they had been in the opposite treatment group. The IRT method is developed to combat existing methods' downfalls such as relying on conditional independence between pre-test scores and outcomes. Using simulation, the IRT method is compared to existing methods under two different model scenarios in terms of Type I and Type II errors. Then the method's parameter recovery is analyzed followed by accuracy of treatment effect evaluation. The IRT method is shown to out perform existing methods in an ability-based scenario. Finally, the IRT method is applied to real data assessing the impact of advanced STEM in high school on a students choice of major, and compared to existing alternative approaches.

Table of Contents

List of Figures

List of Tables

Acknowledgments

Throughout my educational career, I have received immense support from various teachers and friends. Specifically, over the past five years, I have had the joy of being a student of Dr. Matthew Johnson. Dr. Johnson has been a wonderful mentor and I must express my deepest thank you for all of his guidance. His support has made this a wonderful, rewarding, and thought-provoking journey. I would also like to thank my dissertation committee of Young-Sun Lee, Judy Scott-Clayton, Michael Sobel, and Elizabeth Tipton for their support over the past two years as this research idea developed into a completed study.

I would like to thank some of my friends and colleagues for their support during various stages of my research. During data simulation, Michael Chajewski spent many hours listening to my R questions and providing me with expert advice on how to handle certain programming errors. He was always available to listen to me explain my research questions and ideas and then help me turn ideas into programming reality. Over the course of my entire project, Brian Patterson has been a source of great encouragement. Brian has listened to me explain and rehearse my ideas, proof read at various stages, and format, among other supportive tasks. Thank you again for never turning me away when I had a question.

I would also like to thank my best friend, Matthew Korn. He has been by my side for over 11 years of schooling and has always been there to support me. He has played a vital role in my research by helping me proof read my manuscript and also being a willing audience to hear me practice my presentations. Thank you from the bottom of my heart. Thank you to my boyfriend, Sean Meany, for helping out with any support I have needed, no matter how mundane the task.

Finally, I need to thank my mom for all of her support over my 27 years of schooling. Without her loving guidance from day one, none of this would have been possible. Throughout my life as a student, starting at age three, she has been there guiding me, supporting me, and encouraging me even when I might have not believed in myself. Mommy—you have instilled a love of learning in me and a desire to always do my best. For this I am forever grateful.

Dedication

To my parents—on Earth or in Heaven, I will love you forever.

Chapter 1: Introduction

When examining the causal effects of interventions, or experimental factors, statisticians have a number of different methods and data types from which to choose (Shadish, Cook, Campbell, 2001). The preferred method of estimating causal effects is through the design of a controlled experiment. These experiments fully randomize subjects to the levels of interest in the experiment, which implicitly controls for unseen factors that could influence the outcome variable under study; experiments are the gold-standard for investigating causal effects. At the other extreme are observational studies. In observational studies, there is no randomization for selection into treatment and control groups; participants in some sense self-select into the treatment condition. In between these extremes lie variations—studies that have randomization to some level, others that match subjects to try to control for differences, etc.

In many educational settings, true fully-randomized experiments are hard to come by for a number of reasons. One of the most important reasons is that it could be unethical to randomize students into the different intervention groups. For example, if the effects of being retained a year in school were being studied, it would be unethical to randomly assign students to be retained or promoted. A child who truly needs to be held back could potentially end up being promoted, causing him or her to not get the educational help that he or she needs or deserves. While it might be plausible to try out different curricula or remedial programs in an educational setting, certain factors would go into selecting students to participate—maybe only those that need extra help would be placed into one of two experimental remedial courses, this would exclude those who do not need help. This once again goes back to the fact that it is unethical to place students in developmental tracks when they do not need remediation. This same scenario applies to honors courses—it is not ethical to place a student in honors courses when he or she is not ready for that level of rigor.

Non-randomization in studies like those mentioned above leads to a type of selection bias; students are placed into the different treatment levels based on some factor, which is often related to the outcome under study. This type of selection bias can influence the interpretation of the causality of the treatment. If random assignment was used, estimating the causal effect of the treatment is relatively simple and the effectiveness of the treatment or intervention can easily be determined. However, when selection bias is present, standard estimates of the effectiveness of the treatment are invalid; evaluating the treatment or intervention becomes difficult because the potential outcomes are likely related to the selection mechanism itself (Rubin, 1974). That is, selection into the treatment is not independent of the outcome because subjects have not been randomized into the different treatment groups; thus statistical methods based on the assumption of randomization to treatment group (e.g. ANOVA) are not appropriate and would potentially lead to biased results. When selection bias is present in a sample, the researcher must use methods that control for this selection bias, e.g., case-control studies or other matching methods (Rubin, 1974).

Figure 1 depicts a graph describing being placed into treatment solely based on test performance. Imagine a hypothetical college admissions example. Students applying to college must take a placement test to determine if they need to start in a remedial mathematics course. The school would like to see the effect of the remedial course on entering a freshman level course within a timely manner of completing the remedial course. In the figure, X denotes the (potentially multivariate) performance on the placement test, D denotes the decision variable (e.g., to be placed into the remedial mathematics course), and Y_0 and Y_1 denote the potential outcomes (e.g. begin the freshman level course the following semester after completing the remedial course) of students not placed and placed in the remedial course respectively. These potential outcomes are generally expressed by a combination of three factors—treatment placement, observed covariates, and unobserved covariates (Stone, 1993).

According to Item Response Theory (IRT; Lord & Novick, 1968), a test is measure of a true score of an individual, represented by θ in the figure. The potential outcomes and the decision variable, associated with one another through their relationship with test performance, are assumed conditionally independent given test performance. That is that knowing the individual's score on the test will tell you all the information you need to know about the interventions association with the outcome. This is the type of assumption implicitly made by causal inference methods like analysis of covariance (ANCOVA). To properly evaluate the efficacy of the intervention *D,* the statistical method must control for performance on the test, otherwise the estimated effects might be biased (Rosenbaum, 2010). Furthermore, tests have limitations because they cannot measure every possible aspect of an individual and these limitations introduce measurement error into the test scores.

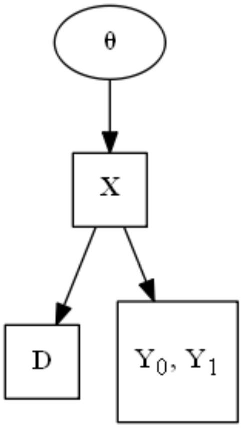

Figure 1: Graph describing the causal inference assumption

There are different ways to control for student differences that can be done before or after treatment. Before placement into treatment, students could be matched on different criteria—gender, age, ethnicity, etc. While not as good at eliminating selection bias as random assignment, this technique can help to limit bias (Rosenbaum, 2010). If matching prior to treatment is not possible, then student differences should be measured. Gender, ethnicity, pre-test scores, and other demographics and relative information should be measured and used to statically control for bias during analysis. The more information collected about the sample, the more likely it is to control for biases using statistical methods. Statistical controls are applied during the data analysis stage and are especially useful for observational data. There have been a number of statistical methods suggested for handling the evaluation of interventions from observational, non-randomized data. They include ANCOVA (Fisher, 1932, as cited in Belin & Normand, 2009), propensity score matching (Rosenbaum & Rubin, 1983), instrumental variables (Wright, 1928; as cited in Angrist, Imbens, & Rubin, 1996), the Heckman Model (Heckman, 1979), and regression discontinuity (Thistlethwaite & Campbell, 1960). However, some of these techniques might be more appropriate in certain situations than in others.

Probably one of the oldest methods for handling data collected as part of such an evaluation is analysis of covariance (ANCOVA; Fisher, 1932, as cited in Belin & Normand, 2009). The typical application of an ANCOVA model, which is appropriate for continuous outcomes, assumes a linear relationship between the outcome (*Y*) and test performance (***X***), and includes a dummy variable for the decision variable (*D*). The effect of *D* is interpreted as the effect of the program. When outcomes are dichotomous, a logistic regression (LR) analysis with a covariate is appropriate in place of ANCOVA. ANCOVA, while simple and popular, has a number of limiting assumptions (e.g., normality, linearity, homogeneity of variance, sample size requirements, homogeneity of relationship between the outcome and test performance).

Potential Outcomes/Counterfactual Model Framework

To begin the discussion of methods for causal inference, the framework used in the potential outcomes model, or counterfactual model, should be set. The counterfactual model has been a part of experimental design and causal inference using observational data for some time, but was formalized by Rubin's work (1974; 1978; 1986). Within this model framework, members of the sample are exposed to a treatment (D = 1) or to a control (D = 0). Each member of the sample has a potential outcome for each treatment condition. Y_{i0} is the potential outcome of individual i under the control and Y_{i1} is the potential outcome of the individual if s/he were to receive the treatment, although only one outcome is observed depending on treatment placement. The observed outcome is

$$Y_i = D_i Y_{i1} + (1 - D_i) Y_{io}$$

The gain from the treatment for an individual, i, is the difference between the individual's two potential outcomes,

$$\Delta_i = Y_{i1} - Y_{io}$$

There are different ways to define the effect of the treatment using this gain, which all depend on the reference population for the effect. These are the average impact of treatment (ATE), average impact of treatment on the treated (TT), and the average impact of treatment on the untreated (UT; Todd, 2006). The following represents these different effects, where H represents a set of characteristics of the individual that are not affected by the treatment. ATE looks at the average treatment effect for all participants—the difference between the treatment and control groups:

$$E(Y_1 - Y_0 \mid H) = E(\Delta \mid H)$$

TT looks at the average treatment effect for those who were treated:

$$E(Y_1 - Y_0 \mid D = 1, H) = E(\Delta \mid D = 1, H)$$

Finally, UT looks at the average treatment effect for those who were not treated:

$$E(Y_1 - Y_0 \mid D = 0, H) = E(\Delta \mid D = 0, H)$$

Pearl's work (2000) can help explain the idea of counterfactuals further and make these terms easier to interpret. A scenario is explained to be counterfactual if unobserved values seem to differ from the observed value. This means that you observe only one outcome, but there are other possible outcomes. However, these outcomes are not observed and hence differ from or contradict each other. Pearl goes on to elaborate that this counterfactual way of thinking can help to influence future decision-making based on causal effects in the past. For example, suppose there is a binary representation, A, of the two academic states a student could be in—struggling (A=1) or not struggling (A=0)—and that there is a binary treatment variable, T representing extra help (T=1) or no extra help (T=0). Now imagine the next two questions a teacher might ask:

Q1: I have a student who is struggling; do I give them extra help?

Q2: I gave extra help to a student who was struggling and now he is no longer struggling. Is it because of the extra help?

Question 1 represents the treatment effect for the student—the difference in student *i*'s potential outcomes. This is represented by $\Delta_i = P(A_{i1} = 0) - P(A_{i0} = 0)$ which is reminiscent of the treatment gain. Question 2 represents a conditional probability statement that is reminiscent of the TT equation above. This is represented as $P(A_{i0} = 1 \mid T_i = 1, A_i = 0)$ and implies the question 'since the student was given extra help and now no longer struggles, what is the probability that the student would still be struggling if no extra help was given?' These relationships are best suited to be interpreted to help make future decisions based on observed casual situations.

According to Rubin, a simple yet strong assumption must be made in using the potential outcomes model for causal inference. This assumption is the stable-unit-treatment-value assumption (SUTVA; Rubin, 1986) and says that one must assume that the outcome of any person exposed to the treatment will remain the same no matter how treatment was assigned or what treatment other people receive. In other words, for N units and T treatments, the value of the outcome for individual i in treatment condition t, Y_{it}, remains the same no matter how individual i was assigned to treatment t and also no matter what treatment the other individuals were exposed to. To even begin discussing causal effects of a relationship, SUTVA needs to be met—the units of study, treatments, and outcomes must be defined to make SUTVA hold. Furthermore, the assumption of strong ignorability (Rosenbaum & Rubin, 1983) must be met as well. Strong ignorability states that given an observed vector of covariates, the potential outcomes are independent of treatment. Mathematically this is represented as

$$(Y_0,Y_1)\perp D \mid H = h \text{, for all } h \text{ and } 0 < \Pr(D = 1 \mid H = h) < 1 \text{ for all } h$$

where h is a vector of observed covariates. This means that controlling for the covariates allows the treatment effect to be estimated without bias.

Modeling Reality

Recall Figure 1 and note that it does not seem to accurately describe reality. For one, a test, X, is a measure of some *true* ability, which is denoted by θ in the figure, and it is likely that potential outcomes, denoted by Y, are related directly to θ rather than being directly related to test performance X. Also, the causal assumptions seen in Figure 1 are $D\perp(Y_0,Y_1)\mid X$, $D\perp\theta\mid X$, and $(Y_0,Y_1)\perp\theta\mid X$ which indicate that the decision, possible outcomes, and ability are pair wise independent given test performance. Recall the college admissions example—it is probably more likely that enrolling in a freshman level course (*Y)* is related to a student's true academic ability (*θ*) rather than how she performed on the admissions placement test. A graph of this scenario is depicted in Figure 2. Here the treatment decision, *D*, and outcomes, *Y*, are conditionally independent given ability, θ*,* which is measured by performance on the test. The causal inference assumptions in this figure are $D\perp(Y_0,Y_1)\mid\theta$, $D\perp(Y_0,Y_1)\mid X$, $D\perp\theta\mid X$, and $(Y_0,Y_1)\perp X\mid\theta$ and the existing methods are still applicable.

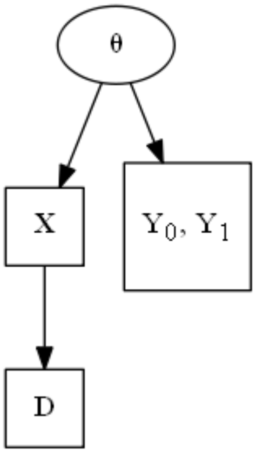

Figure 2: Decision based solely on test performance where the outcome is related to the measured ability rather than test performance itself.

However, what happens when the decision is not directly related to test performance? Imagine another example from college admissions—honors program admission. Suppose a college offers two difficulty levels of honors courses to those students enrolled in the honors program. The college would like to see what effect enrollment in the higher level course has on final cumulative college GPA. It would be feasible to imagine that admittance into the honors program is not based solely on a student's performance on a single placement test, but on multiple factors (placement test included). Furthermore, selection into the difficulty level of the course is up to the student. Here the decision variable, or treatment placement, is no longer directly related to test performance. When the decision variable is not directly affected by performance on a test, but instead related to the true ability, existing methods of analysis would not be appropriate because the strong ignorability assumption of causal inference would be violated—knowing the covariate of test score or performance would no longer mean that treatment placement is independent of the potential outcomes. In mathematical terms $D \perp (Y_0, Y_1) \mid X$ is no longer true. In other words knowing a student's score on a placement assessment would not imply if he elected to take the higher-level honors course. However, both test performance and the decision to take an honors level course are likely to be related to the student's true academic ability (*θ*). Figure 3 describes such a process. In this model the assumptions of causal inference fail since measurement error is no longer captured by treatment placement. This model implies $D \perp (Y_0, Y_1) \mid \theta$, $D \perp X \mid \theta$, and $(Y_0, Y_1) \perp X \mid \theta$.

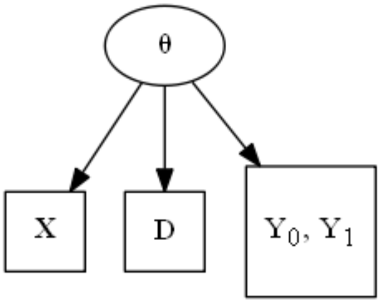

Figure 3: *D* is no longer directly related to test performance, *X,* and causal inference assumptions fail.

Research Goals

The focus of this research is to develop a new technique to overcome the selection bias associated with evaluating the efficacy of an educational program decision or placement (treatment) in an observational setting, specifically when a pre-test is given prior to placement. The existing methods of diminishing selection bias have potential to be lacking the quality to control for innate, unobservable student characteristics as well as test measurement error. However, these characteristics can be captured using Item Response Theory (IRT; Lord & Novick, 1968). Using IRT allows students to be matched based on their innate ability rather than just on observable factors. This research proposes to use student test data, at the item level, to match students (within the IRT model) based on estimates of their true ability and then examine how the treatment differs for students with identical measures of ability. This method of matching is a new and unique approach because previous methods for matching on observable variables can be used in tandem with this method. Comparing those students who possess the same innate ability will allow for selection bias to be controlled for as well as allow causal inferences to be made based on the findings of the observational experiment.

Chapter two of this dissertation begins with a review of the currently available literature on the existing methods of causal inference and their use in educational research. Then, the development of the newly proposed method is described in parallel to the existing methods of causal inference. Chapter three describes the specifics of the model as well as the methods behind the simulation involved in exploring the statistical properties of this new method. In addition, the method is compared to the existing methods to examine the predictive power and added benefit of using the new method. Chapter four provides the results of the simulations and method comparison. Chapter five provides a discussion of this research, limitations, and areas for future work.

Chapter 2: Literature Review

As discussed, there are various types of methods used to evaluate causal inference models. A review of the relevant literature is provided, to understand the use and assumptions of these methods.

Regression Discontinuity

Suppose fifth grade students who are being promoted to the sixth grade take a statewide standardized "fifth grade proficiency assessment" at the end of their fifth grade year. Students who score 50 or below are placed into a developmental track during the sixth grade. In the middle of the sixth grade year, all sixth graders are tested again to see if academic progress is being made. The efficacy of the developmental tracking could be analyzed in this situation using a technique called regression discontinuity.

To truly understand the effect of the developmental track, one would want to compare the difference between the mid-year sixth grade test scores if the student was placed in the developmental track and not placed in the track. However, in reality, only one of these scores is available since the student is either placed in the track or not (Holland, 1986; Rubin, 1974). In potential outcomes framework, there are only two relationships between the test X and the average outcomes, $E[Y_1 \mid X]$ and $E[Y_0 \mid X]$ for those in treatment and not in treatment. Regression discontinuity (RD; Thistlethwaite & Campbell, 1960; Imbens & Lemieux, 2008; Lee & Lemieux, 2009) focuses on those students "right above" and "right below" the pre-test cut point, c. This creates two groups with almost identical scores on the pre-test, but with different treatment placement and therefore outcomes based on different treatment groups. RD then compares the outcomes of these two groups, using regression, and determines the effect of the treatment program, $E[Y_1 - Y_0 \mid X = c \pm \varepsilon]$ as $\lim_{\varepsilon \to 0} c + \varepsilon$, which is $E[Y_1 - Y_0 \mid X = c]$. Further, RD can be extended to cases known as the "fuzzy design" where treatment placement is not strictly determined by some variable—other variables that are unobserved are related to treatment assignment (Hahn, Todd, & van der Klaauw, 2001; Imbens & Lemieux, 2008). The fuzzy design is a special case of the usual, or sharp design, RD analysis. Lee and Card (2008) stress that the classic use of RD, as described above, is suited when the variable assigning treatment is continuous. When it is discrete, they offer corrections that can be used so that RD is still an applicable analysis. Also, the decision variable in RD does not have to be based on only one measure. It can be a created using multiple evaluative measures together and it does not even have to be related to the outcome (Matthews, Peters, & Housand, 2012). However, an RD assumption requires that the assignment be fully known.

RD has been used in a number of educational studies. Recently, Crone, Stoolmiller, Baker, and Fien (2012) examined a multi-component intervention for struggling middle school readers across multiple districts in Oregon using a multilevel cluster RD design. They found that although no significant effect of the intervention was found, there was significant variation in the intervention's effect across the schools in the districts. In two different studies, Abadzi (1984; 1985) examined the effect that ability grouping had on academic achievement and self-esteem. In the first study (Abadzi, 1984), she examined fourth-graders in Texas who were grouped into high and regular ability groups. Abadzi found that those students right above the cut point showed increases in academic performance while those just below the cut point showed decreases after being in a regular ability class for a year. Those in the high ability group showed increases after a year of grouping. There were no significant differences found in self-esteem. In the subsequent study (Abadzi, 1985), she looked at the effects of ability grouping on "long-run" academic achievement and self-esteem of students in Texas in fourth through sixth grade.

The ability-grouping decisions for these students were made at the end of their third grade year and this study examined this grouping up until the end of their fifth grade year. The study found that overall, the students right near the cut point, on either side, were the most affected by ability placement and that the effect from ability-grouping diminished in the long-run for these students. The following studies further illustrate the use of RD in an educational setting.

- Seaver and Quarton (1976) looked at the effect of being placed on the Dean's List for full-time undergraduates in terms of grade point average and course load. They found that earning the distinction of being on the Dean's List encourages students to continue to earn high grades, but does not influence the amount of courses they take on.
- According to Owen (2010), females that earned an A as a final grade in their first economics class had a higher probability of majoring in economics, even after grades were controlled for, indicating that final grades contain valuable feedback that could act as encouragement to pursue further study in a field. Male students did not show the same effect.
- Ou (2010) found that students who barely failed their high school exit exam were more likely to drop out of high school than those that just barely passed the exam.
- Finally, the effect of taking a remedial English class at a community college was explored by Horn, McCoy, Campbell, & Brock (2009). They found that taking the remedial English class had a negative effect on grades in the first level, non-remedial, English class that students needed to take. Furthermore, the longer the time in between the remedial course and the non-remedial course, the greater the negative effect was.

Instrumental Variables

Suppose an elementary school has developed an extracurricular program to help prepare their fifth grade students for transition to the sixth grade, which is housed in the middle school and has many differences from the elementary school. This program is open to all fifth graders and is voluntary to attend. The school wants to see if the program really is beneficial for the transition to the new school. Analyzing the effectiveness of this program just by comparing the sixth grade adjustment of those students who attend to those that did not attend is not appropriate since there are confounding factors that could be associated with attending (i.e. students that adjust well attend easily or students who have trouble adjusting do not attend.) Rather, it would be appropriate to examine the effect of the encouragement of attending the program. To do this, the school has one of the two fifth grade teachers encourage her class to attend the program and tells the other teacher to mention it, but not remind and encourage his students.

However, this is still not a straightforward analysis as there are three student types that would occur. First, there would be students that would go whether encouraged or not encouraged to go. Second, there would be students that would not go whether encouraged or not encouraged. Third, there would be students who would not attend if not encouraged, but would attend if encouraged. It is this group, where encouragement changes the attending decision, that is most important to explore. Furthermore, students in this group of interest might not attend all sessions of the program, in a way creating levels of the treatment. An appropriate technique to analyze a situation like this, usually known as the intent-to-treat (ITT), is instrumental variables. Instrumental variables were originally developed in economics to look at supply and demand (IV; Wright, 1928; as cited in Angrist, Imbens, & Rubin, 1996). IV is appropriate when there is bias associated with the explanatory variable of interested. IV looks at an “instrument,” or variable, which is related to the treatment variable and explains the bias seen the in the explanatory variable of interest. In other words, compliance to attending the program (the instrument) is related to being encouraged to attend (the treatment). IVs are used to control for measurement error in the explanatory variables.

Nomi and Allensworth (2009) used both IV and RD to examine the efficacy of a double-period algebra policy that the Chicago Public Schools put into place. Here, the authors found that about 20% of the students in the study did not adhere to the double-period algebra policy guidelines. Because of this, IVs were used to look at the enrollment effect and a modified RD design was used to look at the policy effect. The lack of complete adherence was confounding the policy effect and needed to be controlled for using IVs. They found that there was a policy effect for those students just above and below the enrollment cutoff score and that there was a positive affect on algebra scores for those that enrolled in the double-period policy than those that were eligible but did not enroll.

Angrist and Lavy (1999) use IV and RD to examine the effect of class size on scholastic achievement in Israeli public schools, where enrollment is the identified instrument, since it is closely related to class size. The study found that overall, there is an increase shown in test scores when class sizes are reduced. This is another illustration of how IV and RD are used together to control for factors that would bias the results.

Propensity Score Matching

In situations where the tracking of students into different tracks is not based on a hard cut point, propensity scores might be used. Propensity score matching creates a matching variable based on the probability to be placed into treatment as a function of multiple observed variables; it is useful to remove biases associated with treatment assignment to estimate treatment effects on an outcome (Rosenbaum & Rubin, 1983). As before, we wish to study the average treatment effect based on the difference in outcomes in the treatment and control groups. Given a selection of observed pretreatment covariates *h*, for any individual *i*, that do not contain all of the observations used to place *i* into a treatment assignment, the propensity score, *e*(*h*), represents the probability to be placed into treatment given the observed covariates.

$$e(h) = pr(D = 1 \mid h)$$

The propensity score is a type of balancing score, *b(h)*, and "is a function of the observed covariates *h* such that the conditional distribution of *h* given *b(h)* is the same for treated (*D* = 1) and control (*D* = 0) units" (Rosenbaum & Rubin, 1983, p 42) and $0 < P(D = 1 \mid b(h)) < 1$. As described in Rosenbaum and Rubin (1983) for large samples, it can be shown that treatment assignment and the observed covariates are conditionally independent given the propensity score,

$$h \perp D \mid e(h)$$

The propensity score is usually unknown in small sample observational studies and must be estimated from the available data. This is done usually using logistic regression and the following equation

$$e(h) = P(D = 1 \mid h) = \frac{P(D = 1)P(h \mid D = 1)}{P(D = 1)P(h \mid D = 1) + P(D = 0)P(h \mid D = 0)}$$

Propensity scores have been used in educational research to study differences between public and private schools (Fan & Nowell, 2011), retention policy for kindergarteners (Hong & Raudenbush, 2005, 2006), the probability of graduating from college if awarded a scholarship (Melguizo, 2010), the effects of completing college on future earnings (Brand & Xie, 2010), and the effects on school choice (Bifulco, 2010).

The Heckman Model

A fourth approach used to estimate a casual effect from observational data is the Heckman model, which uses statistical methods to correct for bias that enters the sample and inference due to nonrandom treatment selection (Heckman, 1979). In educational research, it has been used to correct for biases when analyzing the effect of taking a commercial SAT prep course (Briggs, 2004).

Proposed New method: Ability Matching, Using IRT

The methods described above are all appropriate for processes that fit into the conceptual framework depicted in Figure 1. Some, like ANCOVA and RD, may be able to handle the process described in Figure 2 to estimate the causal effects of an educational decision. These methods must be careful to account for the fact that the potential outcomes are conditionally independent of the decision given true ability, not given test performance. However, these existing methods are not required to provide unbiased estimates of the effects in scenarios depicted by Figure 3.

Recall that in Figure 3, *D* is not conditionally independent of the potential outcomes (*Y*) given test performance, which is a requirement for the standard methods to be appropriate. Because of the conditional relationship between the decision and the outcome, results derived from ANCOVA or propensity score matching are likely to be biased. In fact, as discussed in Rosenbaum (1984) and Holland and Rosenbaum (1986), if test performance $\boldsymbol{X}$, the selection variable *D*, and the potential outcomes Y_0 and Y_1 are all positively associated with the true ability θ, the estimated treatment effects found by matching on test performance $\boldsymbol{X}$ will be positively biased. The same idea follows for the negative association. In a scenario like the one pictured in Figure 3, not only does test performance provide information about a student's true ability, the decision, and outcome also provide more information about the true ability. Providing additional information that does not include measurement error of a test to model the true score allows for better estimates of the true score to be found.

The purpose of this study is to develop a new method for making causal inferences about the effect of a decision on a related outcome when a pre-test is given, in a related domain. In theory, this method should be applicable in both the cut-score scenario, Figure 2, where current methods hold and the ability-only scenario, Figure 3, where current methods struggle with bias. The IRT model implicitly controls for the latent “true” ability and thus allows for direct comparison of treatment and control effects at each level of ability in the domain under study. Thus, the proposed method can be thought of as a sort of model-based matching procedure that matches on an unobservable latent ability.

Using IRT is a new approach to causal inference and will capture additional information about the person through their performance on the test, decision, and outcome and provide a less biased estimate of the effect of the treatment decision. Recall the different college admissions examples, it is possible for two students with different mathematics abilities to achieve the same score on the placement test. For example, questions on tests can be missed by students who have the ability to get the answer correct for many reasons (e.g. distraction in the testing room, fatigue, misreading of the question, etc.). However, IRT models can estimate and use answer patterns and the difficulty of items to gain additional information about an individual’s ability. Using item-level data to gain information about students is like using a fine paintbrush to paint small areas of a canvas. The small brush allows fine grained details to be visible whereas a larger one does not.

For each model scenario (i.e. cut-score scenario and ability-only scenario), this method is evaluated and compared to the applicable existing methods via a simulation study. Then the statistical properties of the method are evaluated using bias and RMSE to check item parameter recovery and treatment effect estimation. Finally, the method is tested on a real world dataset against the other applicable existing methods to comparatively evaluate its performance.

Chapter 3: Method

Item Response Theory Background

To overcome the biases described in the existing methods, this research uses IRT models to examine the causal relationship between a pre-test, educational decision (intervention), and outcome. IRT models are a class of mixed effects, latent variable models for the analysis of repeated ordinal responses. This research assumes that the educational intervention is dichotomous and that the test performance is an ordinal representation of an individual's scores on a battery of test questions. If the outcome variable is also ordinal, then IRT models are appropriate to model situations like that in Figure 2 and Figure 3. According to Mislevy (1991), latent variables not only capture aspects of observable variables, like correctly answering test items, but also capture all associations in various domains, like demographics and aspects of a student's educational standing. Using an IRT model allows the latent ability, θ, to capture unobservable associations during the pre-test, those associated with the decision, and also with the outcome.

IRT contains a broad range of models that allow θ to be estimated by modeling the item responses in a broad range of situations. The simplest model, the Rasch model (Rasch, 1960), sometimes called the one-parameter logistic model (1PL), assumes that the log-odds (logit) of the item characteristic curve is a linear function of θ. The Rasch model only uses the difficulty, β, of each item as a parameter. For item i and examinee j, the probability of j correctly answering i is modeled by the following equation.

$$P_i(\theta_j) = \frac{e^{(\theta_j - \beta_i)}}{1 + e^{(\theta_j - \beta_i)}}$$

However, in many situations the Rasch model is too restrictive and is generalized to the two-parameter logistic model (2PL; Birnbaum 1968), which allows the slope of the log-odds to vary. In the 2PL model there are two parameters for each item—the item difficulty, *β*, and the item discrimination, *α*.

$$P_i(\theta_j) = \frac{e^{\alpha_i(\theta_j - \beta_i)}}{1 + e^{\alpha_i(\theta_j - \beta_i)}}$$

The 2PL model can be generalized even further to account for guessing, *c*, on questions, because in many situations when someone does not know the answer to a question, they guess rather than just getting it incorrect. This generalization is known as the three-parameter logistic model (3PL; Birnbaum, 1968).

$$P_i(\theta_j) = c_i + (1 - c_i)\frac{e^{\alpha_i(\theta_j - \beta_i)}}{1 + e^{\alpha_i(\theta_j - \beta_i)}}$$

These three models (Rasch, 2PL, and 3PL) are the most common models for dichotomous item response data. There are specific models for data that is polytomous, such as the generalized partial credit model (GPCM; Muraki, 1992) for ordinal polytomous responses, and the nominal response model (NRM; Bock, 1972) for nominal item responses. These polytomous models, estimate the probability of scoring in category *l* of a *k*-category model; for the GPCM this probability is conditional on the fact that the score is either in category *l*-1 or *l*. The GPCM is given by

$$\frac{P_{ijl}}{P_{ij,l-1} + P_{ijl}} = \frac{e^{\alpha_i(\theta_j - \beta_{il})}}{1 + e^{\alpha_i(\theta_j - \beta_{il})}}$$

The NRM is given by

$$P_{il}(\theta) = \frac{e^{(\alpha_{il}\theta_j + c_{il})}}{\sum_{g=1}^{k} e^{(\alpha_{ig}\theta_j + c_{ig})}}$$

Another advantage is that all of these models can be modeled using standard software like PARSCALE (Muraki & Bock, 1997), BILOG-MG 3 (Zimowski, Muraki, Mislevy, & Bock, n.d.), and packages in R like LTM (Rizopoulos, 2011) and IRTOYS (Partchev, 2012).

When item responses are ordinal, the decision variable is dichotomous, and the potential outcomes are ordinal, the conceptual models depicted in Figure 2 and Figure 3, can be thought of as generalized linear or non-linear mixed effects models. Typically these models assume that θ, which represents ability, is distributed as a normal random variable with a mean of zero and variance of one, $N(0,1)$, and an IRT model can be applied to the data. The path from θ to $\boldsymbol{X}$ represents the relationship of test items and ability. For dichotomously scored items, standard IRT models (Rasch, 2PL, and 3PL) can be used to model the data. If the items are polytomously scored, like partial credit items, a GPCM is a plausible model for analyzing this relationship and expanding the applications of this model. Mathematically, the GPCM will reduce to a standard IRT model if the items are scored dichotomously; the polytomous version of the 2PL model is the GPCM. The path from θ to D represents the probability of being placed into the treatment condition based on ability. Simplest is dichotomous, but it can also be extended to polytomous (nominal and ordinal) conditions. This relationship can be modeled using the standard IRT models. Depending on the scenario examined, Figure 2 or Figure 3, will determine if this path is included in the overall model. The paths from θ to Y_0 and θ to Y_1 represent the probability of success on the outcome variable in each treatment group, given the student's ability. For these models the outcome is represented in ordered categories, either dichotomously (e.g. success or failure) or polytomously (e.g. low, medium, high), but not continuously. A student only has an observed value for one of these paths since only one decision can be followed. The value would be missing for the other path. This relationship is also modeled with standard IRT models and can be extended from the dichotomous condition to polytomous (nominal and ordinal) conditions.

Under the IRT model the outcome variables, Y_0 and Y_1 are treated as "items" in a large "test." Since they are treated as "items" they contain the "item responses" for the individuals in the sample. Certain individuals have missing data for these responses. As long as these outcomes represent situations under the cut-score scenario (Figure 2) or the ability-based scenario (Figure 3) the data would be considered missing at random (MAR; Rubin, 1976). The data is considered MAR since it is impossible for an individual to have an outcome for both the treatment and control group.

The overall IRT model links all of these paths together as if it was one large set of item responses where the test items, $\boldsymbol{X}$, come first, followed by the decision D, and the outcome variables Y_0 and Y_1. Representing the data as such allows for the GPCM, in general terms, to be applied to this "large test" and parameter estimates can be estimated for each outcome path. These estimates represent the discrimination and difficulty of success on the outcome for each treatment group—the outcomes for the treatment and control groups are like items on the test and can be compared to each other as such.

Data Simulation

Simulation Conditions To compare this method to existing causal inference methods and to study the statistical properties of this method, simulation conditions were formed from combinations of sample size, pre-test length, and true treatment effect. Table 1 lists the possible values that each aspect of the simulation could contain.

Sample sizes ranged from 500 to 10,000 and were chosen to realistically model different groups of test takers—from a placement examination of an incoming college freshman class to a sample of students taking an admissions test, like the SAT. The number of items on the pre-test could take on three different values of 10, 20, or 50, also chosen to represent true life testing situations such as a college placement test or an admissions exam (College Board, 2013).

The difficulty parameters of the outcome variables were chosen to represent the difference between the two groups. The difference in the difficulty parameters β_1 and β_0 is a measure of the treatment effect; it represents the difference in log-odds of a student at the mean ability level, $\theta = 0$. In very general terms, a treatment effect either exists or it does not exist. If an effect exists, it can be either positive or negative as well as either weak or strong. To replicate these options, five different treatment effect representations were used. For all of the outcome conditions, the discrimination parameter of the outcome α, (i.e. the slope) was set constant at 1. This allowed the representations of possible treatment effects to only be represented by the difficulty parameters. It is possible to have identical outcome difficulties, but varying slopes, which would result in differential effects.

No effect would exist if both treatment and control groups found the outcome equally difficult, as long as they had the same outcome slopes. Finding the outcome equally difficult implies that there is no difference in success between those in treatment and those in the control group with the same level of ability, indicating no treatment effect, represented by the "No Difference" condition in this study. A negative effect exists if the control group performs better on the outcome than the treatment group. When the treatment group finds the outcome to be more difficult than the control group for those with the same ability, the treatment group will have a lower probability of success on the outcome. This effect can be either weak or strong, represented here by the negative low (Low -) and negative high (High -) conditions. A positive effect exists if, conditional on having the same ability, those in the treatment group perform better on the outcome than those in the control group. This means those in the treatment group find the outcome less difficult than those in the control group. The positive effect can also be weak or strong, represented by the positive low (Low +) and positive high (High +) condition.

Table 1

Values for simulation conditions.

Sample Size (N)	Number of Items (X)	Treatment Effect	Y_0 (Control)	Y_1 (Treatment)
500	10	No Difference	0	0
1000	20	Low -	-0.5	0.5
5,000	50	Low +	0.5	-0.5
10,000		High -	-1	1
		High +	1	-1

The design of the simulations creates all possible combinations of these values, resulting in 60 different simulation conditions. For example, a simulation of 500 students on a 10-item test with item difficulties for the outcomes of 0 and 0 for the control and treatment groups respectively is created. Followed by a simulation of 500 students on a 10-item test with item difficulties for the outcomes of -0.5 and 0.5 for control and treatment groups respectively were created, etc. For the simulation conditions involving a cut score, the cut score value was set to be equal to half of the item on the test. This resulted in a split of approximately 70% of the sample being placed into treatment, which is a realistic value according to the literature (Legislative Analyst's Office, 2011).

Pre-test True Item Parameters The initial difficulty and discrimination values for the 10, 20, and 50 items on the test were modeled based on true parameters from the 2007 Trends in

Mathematics and Science Study (TIMSS), see Table A1 in the Appendix for these values. The TIMSS is an international study that compares the science and mathematic achievement of U.S. students to that of students in other countries. The mathematics items are modeled by a 3PL IRT model (Gonzales, Williams, Jocelyn, Roey, Kastberg, & Brenwald, 2008). Since the current research uses a 2PL model, rather than a 3PL model, only the difficulty and discrimination parameters of the TIMSS items were selected. For each item parameter (discrimination and difficulty), the empirical density of the parameters used in the TIMSS were plotted and a normal curve fitted to the density. Then, the parameters of the normal curve were used to sample 50 items each for difficulty and discrimination to be used as the initial values for the simulation pre-test. These were then broken down into each test length, with each longer test containing the smaller tests. The first ten items represent the 10-item test; the first 20 items represent the 20-item test and contains the 10-item test within it. Finally, the 50-item test is represented by all 50 items, which contain the other two tests. These values are considered the true parameter values.

Dataset Building Each of the 60 simulation conditions contained 1,000 replications of data simulation. Each replication consisted of sampling “true” ability values from a standard normal distribution (N(0,1)) for the appropriate sample size. For each member of the sample size, a 2PL model was used to estimate dichotomous item responses on the appropriate number of pre-test items using the initial item parameter values. Using the specific outcome difficulty for each simulation condition (e.g. “High +”), a 2PL model was used to generate dichotomous responses on each outcome (i.e. the outcome for the treatment group and the outcome for the control group), based on each member’s ability. Then, dichotomous treatment placement was also estimated using a 2PL model, the decision parameters (difficulty = 0, discrimination = 1), and each member’s ability. This decision is called the ability-based decision. A second treatment decision was also assigned to the member based on the sum of that member's correct responses' relationship to an established cut score. The cut score was designated as 50% accuracy on the test. If the member's test score was equal to this value or less, that member was assigned to the treatment group. This decision variable is referred to as the cut score decision. These two decisions variables could result in identical or different treatment placement.

Recall that a member can only have a value of the outcome for the treatment group in which that member was assigned. Therefore, two sets of outcome variables were created that corresponded to the ability-based decision and the cut score-based decision. If a member was assigned to the treatment group, that member's value in the control group's outcome variable was set to missing, and vice versa. This was done for both decision variables. An example of the dataset set up can be seen in Table 2.

Table 2

Sample simulated dataset.

Member	θ	I_1	I_2	.	.	.	I_X	TS	D_θ	y_c	y_t	D_{cut}	y_{c_cut}	y_{t_cut}
1	0.585	0	1	.	.	.	0	1	0	1	NA	1	NA	0
2	-0.477	1	1	.	.	.	0	2	1	NA	1	1	NA	0
3	0.902	0	0	.	.	.	0	0	1	NA	0	0	0	NA
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
N	-1.626	1	0	.	.	.	0	7	0	0	NA	0	1	NA

Note: I_1 to I_X represents each item on the pre-test; N represents sample size, TS represents the total score on the pre-test; D_θ represents the ability-based decision with corresponding outcomes of y_t and y_c; D_{cut} represents the cut score-based decision with corresponding outcomes of y_{t_cut} and y_{c_cut}

Causal Inference Method Comparison

Recall the cut-score scenario in Figure 2 and the ability-based scenario in Figure 3. These figures represent two different representations of how the pre-test, decision, and outcomes can be associated. As explained earlier, not all causal inference methods are applicable to both scenarios. However, the IRT causal inference method must be evaluated in each scenario and compared to the appropriate methods to determine the best use of the method.

Cut-score Scenario Three different causal inference methods are compared within the cut-score scenario, shown in Figure 2. They LR, RD, and the IRT method. Each method requires different combinations of each simulated data set, but all use the cut-score based decision (D_{cut}) as well as the cut-score associated outcomes (y_{t_cut} and y_{c_cut}).

Logistic regression. Since the outcome in this study is dichotomous a logistic regression model (LR) is used in place of an ANCOVA model. This LR model requires three pieces—a covariate, a selection variable, and an outcome. For this model the covariate is the total score on the test and is centered around the treatment group's mean. The selection variable representing treatment placement is the cut score-based decision variable. Combining both treatment group outcomes into one variable creates a single outcome variable. Since each member has an outcome value either in the treatment condition or the control condition there are no missing entries in this composite variable. A logistic regression using the binomial distribution with a logit link function was used to estimate the treatment effect. This model returns estimates of the treatment effect, the interaction between total score and treatment placement, and associated significance values.

Regression discontinuity. The RD method required similar pieces as the LR model, except that the total score was centered around the cut score, rather than the treatment group's mean score. A logistic regression using the binomial distribution with a logit link function was also used to estimate the treatment effect. This model returns estimates of the treatment effect, the interaction between total score and treatment placement, and associated significance values.

IRT-based approach. A 2PL model was used to estimate ability based on item responses and each outcome. The individual pre-test items were entered into the model using dichotomous scoring (correct/incorrect) and the outcome variable for each treatment group was also entered dichotomously (success on the outcome/non-success) as an item. Together this created one large "test" that consisted of 12, 22, or 52 items. The estimation was done in R using the *est( )* function within the package IRTOYS (Partchev, 2012). Specifically, this function used the ICL (Hanson, 2002) estimation program to estimate the 2PL model. This method returns difficulty and discrimination parameter estimates for each "item" in the model.

Ability-based Scenario Three different causal inference methods are compared within the ability-based scenario, shown in Figure 3. They are LR, propensity score matching, and the IRT method. Each method requires different combinations of each simulated data set, but all use the ability-based decision (D_θ) as well as the associated outcomes (y_t and y_c).

Logistic regression. Once again the LR model requires three pieces—a covariate, a selection variable, and an outcome. For this model the covariate is the total score on the test and is centered around the treatment group's mean. The selection variable representing treatment placement is the ability-based decision variable. The outcome variable is created by combining both treatment group outcomes into one variable. Since each member has an outcome value either in the treatment condition or the control condition there are no missing entries in this variable. A logistic regression using the binomial distribution with a logit link function was used to estimate the treatment effect. This model returns estimates of the treatment effect, the interaction between total score and treatment placement, and associated significance values.

Propensity score matching. A propensity score was created using the test items as the observed covariates. Once the propensity score was created, the *Matching* ( ) function in R was used to match those in treatment and control, based on the propensity score. This matching was executed as one-to-one matching with replacement. This model then estimates the probability of success on the outcome using the matched sample and returns an estimate of the treatment effect.

IRT-based approach. A 2PL model was used to estimate ability based on item responses, the decision, and outcome. The individual pre-test items were entered into the model using dichotomous scoring (correct/incorrect), the ability-based decision variable was also entered into the model dichotomously, as was the outcome variable for each treatment group. Together this created one large “test” that consisted of 13, 23, or 53 items. The estimation was done in R using the *est( )* function within the package IRTOYS. Specifically, this function used the ICL estimation program to estimate the 2PL model. This method returns difficulty and discrimination parameter estimates for each “item” in the model.

Parameter Recovery Evaluation Parameter estimates from each model were compared to the true values of the parameters in terms of biases and root mean squared errors (RMSE) to evaluate parameter recovery. The bias of the estimate of the parameter of interest, γ, is calculated as the difference between the mean estimated parameter (over all replications) and the true parameter value, or

$$BIAS(\hat{\gamma}) = E[\hat{\gamma}] - \gamma$$

The RMSE of the estimate of the parameter of interest is calculated as the square root of the sum of the squared bias of the estimated parameter and the variance of the estimated parameter.

$$RMSE(\hat{\gamma}) = \sqrt{(BIAS(\hat{\gamma}))^2 + \mathrm{var}(\hat{\gamma})}$$

The treatment effect was calculated for each method within a scenario. These estimates were compared to the true effect in terms of bias and RMSE. Then, these statistics were compared between models.

Method Comparison Statistics Within each scenario, the three models were compared to each other in terms of Type I and Type II errors. For each causal inference method, the proportion of simulated datasets where statistically significant effects, $p \leq 0.05$, were found was calculated. When no true differences exist in the simulated data (i.e. the No Difference group), this proportion is an estimate of the Type-I error rate. When there are true differences (i.e. Low -, Low +, High -, High +), this proportion is an estimate of the power, or one minus the Type-II error rate.

For the IRT models, the parameter estimates for the difficulty from treatment and control conditions for each simulation replication, *i*, were used with the MSE (the square of the RMSE) and bias for the overall condition, *j*, to create a z-score. This z-score was then compared to a normal distribution to get a resulting *p*-value.

$$z = \left| \frac{\hat{\beta}_{t\_ij} - \hat{\beta}_{c\_ij}}{\sqrt{MSE_j - (Bias_j)^2}} \right|$$

For the propensity score model the *p*-value was calculated by comparing a test statistic to a *t*-distribution. The test statistic was calculated using the estimated effect and the standard error with degrees of freedom equal to one less than the number of observations from each simulation condition.

$$t = \left|\frac{est}{se}\right|$$

Real World Data

To examine this method on real world data, data from a large-scale college admission test as well as data from an end-of-year high school exam were obtained. The main research question using this data is to look at the effect of taking an advanced course in one of the science, technology, engineering, and mathematics (STEM) areas during the senior year of high school and its association with majoring in a STEM field in college. Data from the mathematics section of the admissions test, taken during the junior year, acts as the pre-test. The decision is if the student took the advanced STEM course during their senior year. The outcome is majoring in a STEM field in college. There is an interest in the STEM fields, and it seems particularly useful to identify aspects of high school curriculum that could help to interest students in STEM areas.

Chapter 4: Results

This chapter discusses the results of this dissertation. The two different scenarios—cut-score (Figure 2) and ability-based (Figure 3)—will be examined separately. The investigation of the causal inference methods under each scenario will begin with the presentation of the estimated item parameters, a discussion of the Type I and Type II error rates, and a discussion about model parameter recovery and treatment estimation. The individual causal inference methods are compared briefly, however Chapter 5 will present a thorough comparison and discussion. This section begins with the ability-based scenario (Figure 3), moves to the cut-score (Figure 2), and finishes with an analysis of real world data.

The *est( )* function in the IRTOYS (Partchev, 2012) package in R was used to estimate the 2PL models by calling the ICL program (Hanson, 2002). It was observed that multiple fitting errors in ICL occurred for the 2PL model under the cut-score scenario. To overcome this, a warning and error catching function was written so that a new dataset was simulated whenever a fitting error occurred. The dataset that triggered the error or warning was saved for future investigation, but not included in the model running process.

Comparison Under the Ability-based Scenario

Recall that the ability-based scenario is the model where the decision is not based on the pre-test, but is based on the student's ability (see Figure 3). In this scenario, the student is placed into treatment based on his/her probability of success on the decision, which is based on the student's ability. The mean parameter estimates for the outcome variables in the ability-based scenario, along with the true values, RMSE, and bias can be seen in Table A2 through Table A5 in the Appendix.

Type I and Type II Error Rates The main hypothesis of this study is to determine how the IRT based method compares to existing methods of causal inference. For the three models—LR, propensity score, and IRT—the proportion of simulated datasets where statistically significant effects, $p \leq 0.05$, were found was calculated for each simulation condition. Table 3 through Table 6 show the proportion of significant differences found by each method for the outcome conditions. For all simulations involving the No Difference outcome condition, the IRT method found a proportion of about 0.05, indicating that it is performing as expected. This value remained relatively constant as the sample size and test length varied. The other methods, specifically LR and propensity score matching, have high Type I error rates as shown by the high proportion of significant differences found for this condition. For both LR and propensity score the Type I error decreases as test length increases and yet increases as sample size increases.

The conditions where true differences exist (i.e. Low -, Low +, High -, High +) represent the power of the test. The higher the power, the better the test is at detecting true differences. A good test is a test that has a power function near one for most values of the estimator that fall in the rejection region (Casella & Berger, 2002, p. 383). In terms of power, the IRT method demonstrates higher power than propensity score matching and similar power to LR. This is shown by the fact that a higher proportion of significant differences are found by the IRT method than in propensity score matching, within a simulation condition. Also, for all methods in almost all conditions, as test length and sample sized increased independently power also increased.

Together this information indicates that the IRT method is able to detect differences between the treatment and control groups better than the existing methods for the ability-based scenario, when these differences exist. Also, the IRT method is less likely than the other methods to detect difference when no difference between treatment and control exists.

Table 3

Proportion of statistically significant effects found by each model, under the ability-based scenario, for a sample size of 500.

Simulation Characteristics		**IRT**	**LR**		**Propensity Score**
X	**OC**		**Decision**	**Interaction**	
10	**No Difference**	0.052	0.329	0.063	0.216
	Low -	0.995	0.822	0.066	0.590
	Low +	0.655	1.000	0.043	0.986
	High -	1.000	1.000	0.061	0.999
	High +	0.843	1.000	0.051	1.000
20	**No Difference**	0.048	0.101	0.060	0.093
	Low -	0.991	0.934	0.086	0.705
	Low +	0.872	0.997	0.062	0.946
	High -	1.000	1.000	0.076	0.999
	High +	1.000	1.000	0.043	1.000
50	**No Difference**	0.052	0.060	0.064	0.036
	Low -	1.000	0.975	0.086	0.712
	Low +	0.930	0.990	0.059	0.783
	High -	1.000	1.000	0.111	0.996
	High +	1.000	1.000	0.051	0.999

Note: X represents the test length and OC represents the outcome condition.

Table 4

Proportion of statistically significant effects found by each model, under the ability-based scenario, for a sample size of 1,000.

Simulation Characteristics		**IRT**	**LR**		**Propensity Score**
X	**OC**		**Decision**	**Interaction**	
10	**No Difference**	0.048	0.583	0.058	0.379
	Low -	1.000	0.985	0.063	0.904
	Low +	0.991	1.000	0.054	1.000
	High -	1.000	1.000	0.063	1.000
	High +	1.000	1.000	0.055	1.000
20	**No Difference**	0.052	0.147	0.075	0.117
	Low -	1.000	1.000	0.113	0.966
	Low +	0.998	1.000	0.058	1.000
	High -	1.000	1.000	0.127	1.000
	High +	1.000	1.000	0.055	1.000
50	**No Difference**	0.047	0.063	0.071	0.052
	Low -	1.000	1.000	0.115	0.965
	Low +	1.000	1.000	0.073	0.994
	High -	1.000	1.000	0.146	1.000
	High +	1.000	1.000	0.050	1.000

Note: X represents the test length and OC represents the outcome condition.

Table 5

Proportion of statistically significant effects found by each model, under the ability-based scenario, for a sample size of 5,000.

Simulation Characteristics		**IRT**	**LR**		**Propensity Score**
X	**OC**		**Decision**	**Interaction**	
10	**No Difference**	0.047	0.998	0.095	0.970
	Low -	1.000	1.000	0.118	1.000
	Low +	1.000	1.000	0.058	1.000
	High -	1.000	1.000	0.149	1.000
	High +	1.000	1.000	0.056	1.000
20	**No Difference**	0.050	0.519	0.228	0.470
	Low -	1.000	1.000	0.417	1.000
	Low +	1.000	1.000	0.139	1.000
	High -	1.000	1.000	0.499	1.000
	High +	1.000	1.000	0.075	1.000
50	**No Difference**	0.042	0.082	0.257	0.108
	Low -	1.000	1.000	0.409	1.000
	Low +	1.000	1.000	0.118	1.000
	High -	1.000	1.000	0.540	1.000
	High +	1.000	1.000	0.080	1.000

Note: X represents the test length and OC represents the outcome condition.

Table 6

Proportion of statistically significant effects found by each model, under the ability-based scenario, for a sample size of 10,000.

Simulation Characteristics		**IRT**	**LR**		**Propensity Score**
X	**OC**		**Decision**	**Interaction**	
10	**No Difference**	0.059	1.000	0.136	1.000
	Low -	1.000	1.000	0.176	1.000
	Low +	1.000	1.000	0.082	1.000
	High -	1.000	1.000	0.232	1.000
	High +	1.000	1.000	0.061	1.000
20	**No Difference**	0.047	0.815	0.422	0.835
	Low -	1.000	1.000	0.651	1.000
	Low +	1.000	1.000	0.291	1.000
	High -	1.000	1.000	0.789	1.000
	High +	1.000	1.000	0.088	1.000
50	**No Difference**	0.047	0.143	0.446	0.222
	Low -	1.000	1.000	0.690	1.000
	Low +	1.000	1.000	0.217	1.000
	High -	1.000	1.000	0.842	1.000
	High +	1.000	1.000	0.065	1.000

Note: X represents the test length and OC represents the outcome condition.

To visualize this further the proportion of statistically significant effects from each condition of the three methods were plotted in a histogram by each treatment effect outcome group. The No Difference condition is shown in Figure 4 through Figure 6 for the various methods. Figure 4 shows the No Difference outcome condition for the IRT method. According to Casella and Berger (2002, p. 397-8), *p*-values that fall under the null hypothesis are uniformly distributed. Since this outcome condition represents the null hypothesis (i.e. the situation where both treatment and control group find the outcome equally difficult), these histograms should look like a uniform distribution. As you can see from Figure 4, this is true for the IRT method. However, the LR method (Figure 5) and the propensity score method (Figure 6) do not resemble a uniform distribution. The remaining plots for the other outcome conditions can be seen in the Appendix—Figure A1 through Figure A4 for the IRT method, Figure A5 through Figure A8 for the LR method, and Figure A9 through Figure A12 for the propensity score method.

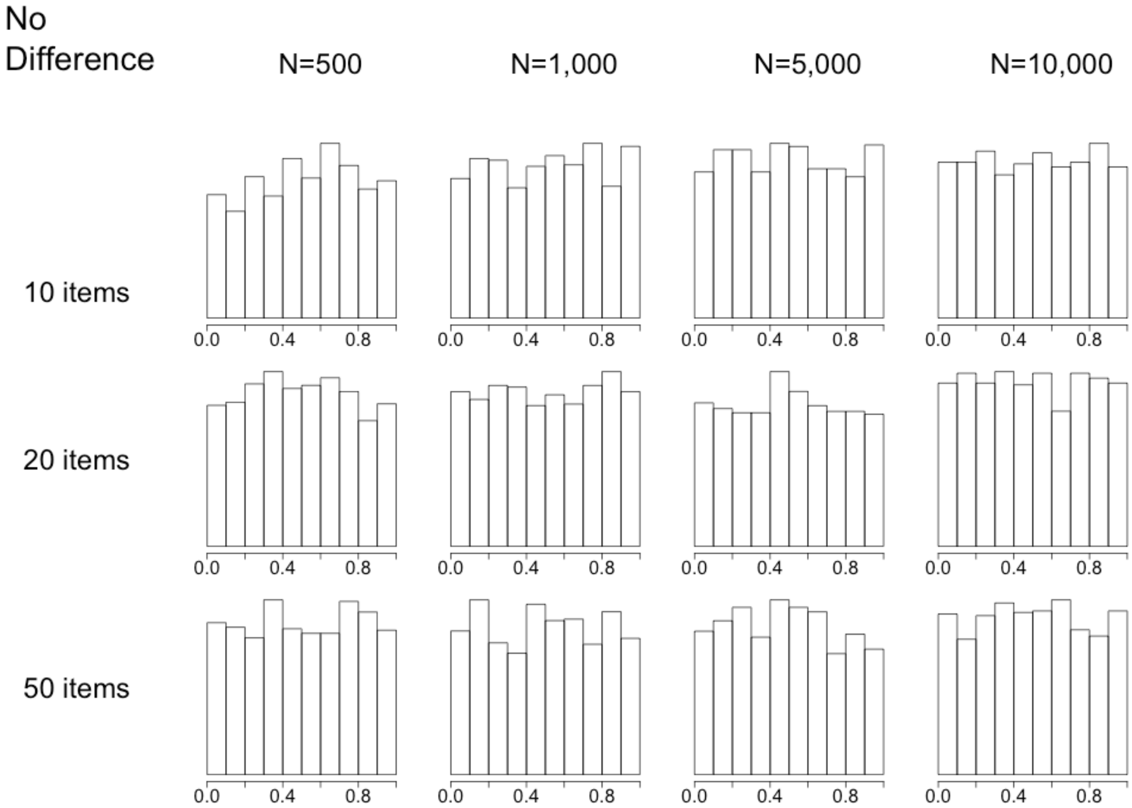

Figure 4: Histograms of the *p*-values from the IRT method under the ability-based scenario for the No Difference outcome group.

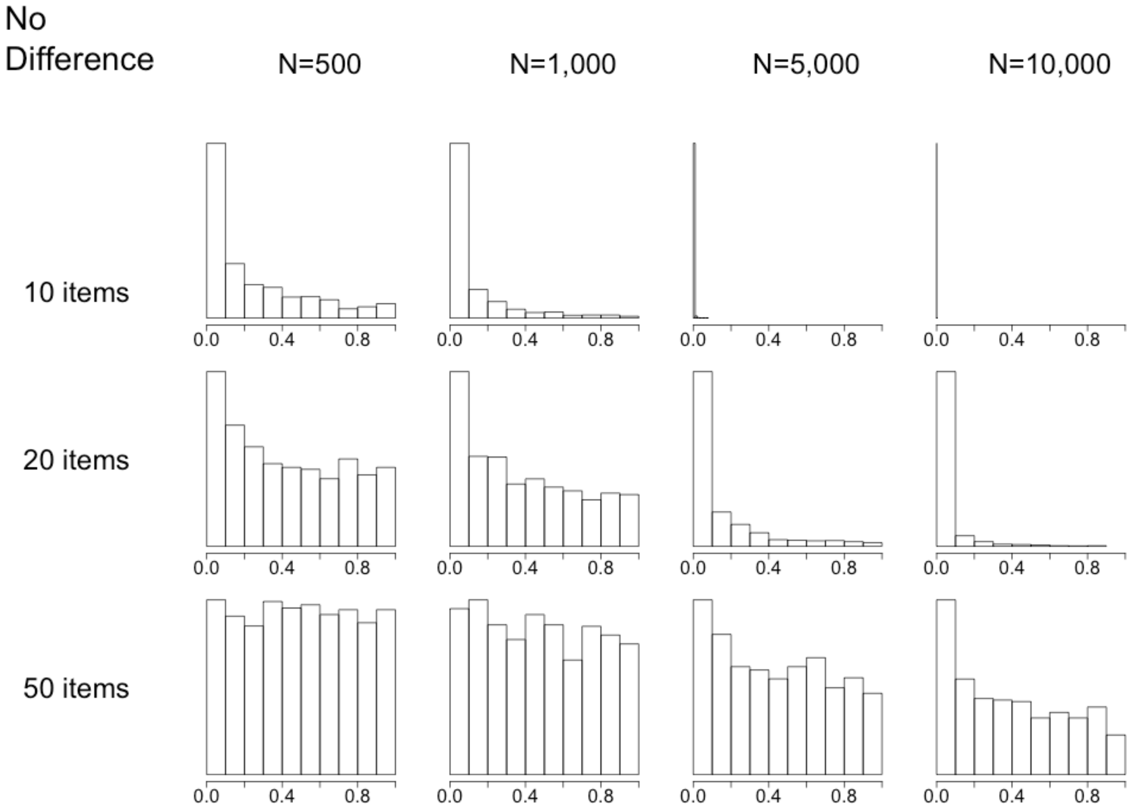

Figure 5: Histograms of the *p*-values from the LR method under the ability-based scenario for the No Difference outcome group.

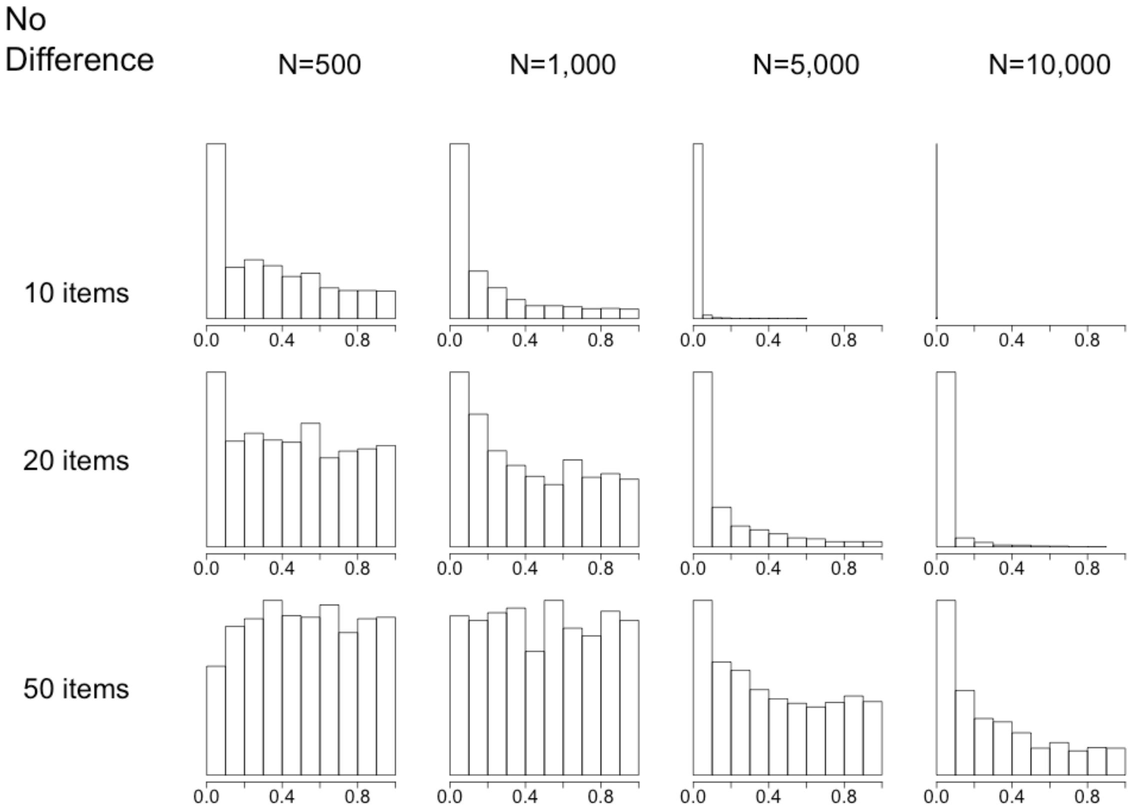

Figure 6: Histograms of the *p*-values from the propensity score method under the ability-based scenario for the No Difference outcome group.

Bias and RMSE for the IRT Method The next step in the investigation of the IRT method was to examine parameter recovery within the confines of the ability-based scenario. To do this, mean biases and RMSE were calculated for the discrimination and difficulty parameters estimated by the IRT method. To examine how these values change between the different aspects of each simulation condition, profile plots were created for combinations of each treatment group's outcome and the associated item parameter. Figure 7 and Figure 8 show plots for the difficulty parameter of the treatment group's outcome under the ability-based scenario.

Figure 7 shows the profile plot for the bias of the difficulty parameter of the outcome variable of the treatment group. The five panels represent each outcome condition, No Difference through High +, with the lines representing the test length and the x-axis showing the sample size. A common occurrence in each of these panels is that the bias decreases as sample size increases. In addition, in the smallest sample size condition, N=500, the 10-item test shows the highest bias and the 50-item test the smallest. This same pattern can be observed for the bias of the difficulty in the control condition (see Figure A13 in the Appendix) as well as for the bias of the discrimination parameters in both groups (see Figure A14 for the treatment group and Figure A15 for the control group in the Appendix).

Figure 8 shows the RMSE of the difficulty parameter of the outcome variable in the treatment group. The panels in this plot show the same trend as those in Figure 7 and the associated plots in the Appendix. As the sample size increases, the RMSE gets smaller and approaches zero. In other words, as sample size increases parameter recovery gets better. In addition, it is interesting to note that as test length becomes longer, RMSE is smaller. This is specifically shown in the smallest sample size condition, but disappears as the sample size increases. This same trend is also shown in the plots of the RMSE of difficulty parameter for the outcome variable in the control group (Figure A16 in the Appendix) as well as for the RMSE of the discrimination parameter of the outcome variables in both treatment (Figure A17 in Appendix) and control (Figure A18 in Appendix) groups. For completeness, scatterplots of the bias and RMSE for all of the remaining test items, including the decision variable, can be seen in the Appendix. For the ability-based scenario see Figure A19 to Figure A24 for the bias and Figure A25 to Figure A30 for the RMSE.

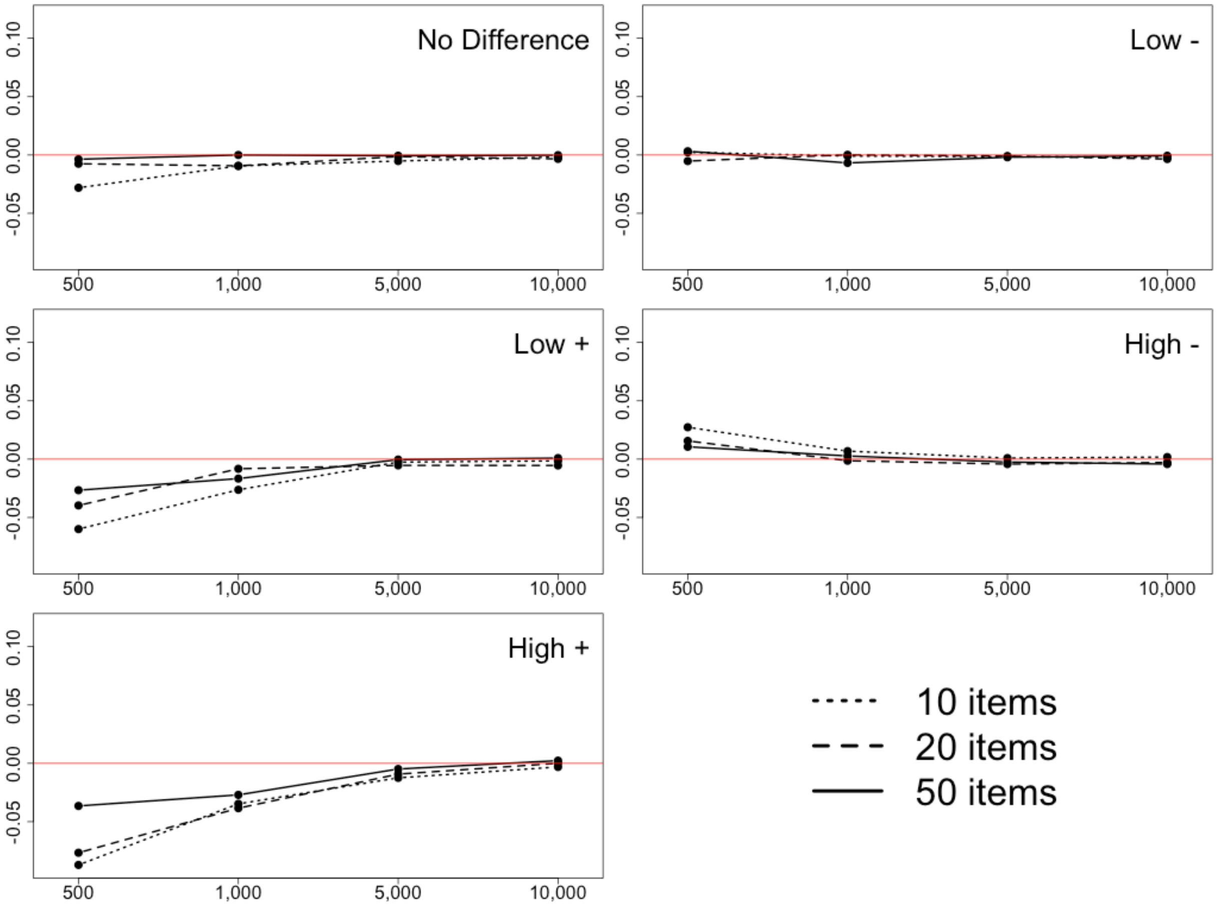

Figure 7: Profile plot for the bias of the difficulty parameter of the outcome variable for the treatment group under the ability-based scenario.

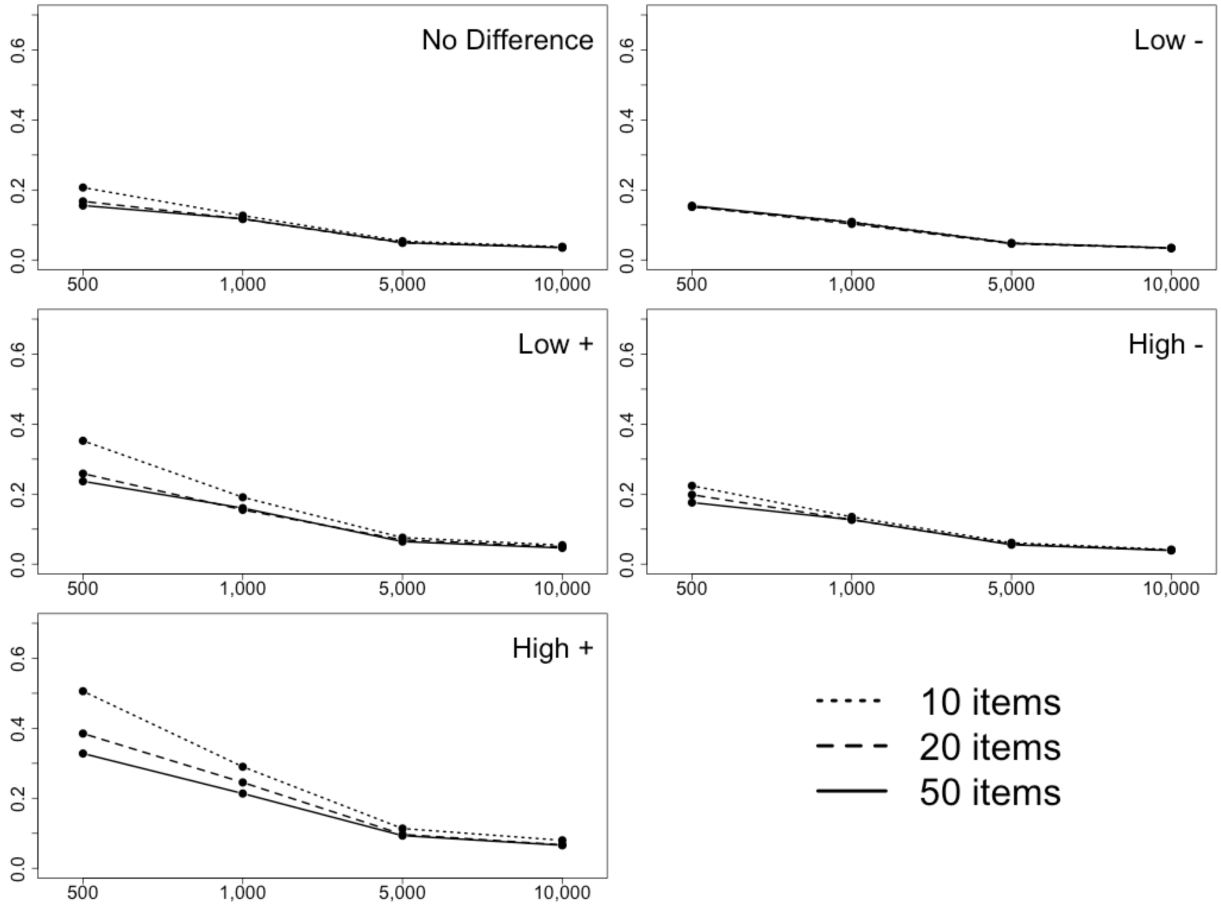

Figure 8: Profile plot for the RMSE of the difficulty parameter of the outcome variable for the treatment group under the ability-based scenario.

ANOVA and ANCOVA Analysis of Simulation Conditions From the previous profile plots it can be seen that the bias and RMSE values vary between the different simulation conditions in terms of sample size, pre-test length, and outcome difficulty condition, indicating that further investigation is needed. To do this, individual three-way factorial design ANCOVA models were run to see if certain components of each condition—sample size, pre-test length, outcome condition (No Difference, Low -, Low +, High -, High +)—explained the variation in bias and RMSE values while controlling for parameter estimates. These three-way factorial models had equal groups in each cell of the design, making effect interpretation easier. Five models used a corresponding bias as an outcome and fived used a corresponding RMSE. For example, one ANCOVA used the RMSE of the difficulty parameter for the treatment group, controlled for the estimate of the difficulty of the treatment group, and looked at sample size, pre-test length, and outcome condition.

ANCOVA models with full interactions were estimated, but had “an essentially perfect fit” according to R and were unreliable. Therefore, the interactions were removed and only main effect models were fit. For the bias, these main effect ANCOVA models had the same “essentially perfect fit” and therefore the parameter estimate portion of the model (difficulty, discrimination, or difference in difficulty) was removed from each model and an ANOVA model was fit instead. The results of these ANOVA and ANCOVA analyses are shown in Table 7 and Table 8. Table 7 shows the association with bias. Sample size is always significantly associated with the bias, regardless of the parameter being estimated. The outcome condition is significantly associated with bias only for the difficulty parameter. Pre-test length was significantly associated with the bias in the difficulty parameter for the treatment group and in the difference of the difficulty between the treatment and control groups. Table 8 shows the results of the RMSE analysis. Once again sample size is always a significant factor. Outcome condition is significantly related in all but the RMSE of the discrimination parameter of the control group. Yet, pre-test length is rarely significantly associated.

Table 7

ANOVA table for the bias in the ability-based scenario.

Response	Source	DF	F
Difficulty **y_c**	Outcome Condition	4	7.319*
	Sample Size	3	3.435*
	Pre-test Length	2	1.928
	Residual	50	
Difficulty **y_t**	Outcome Condition	4	8.275*
	Sample Size	3	5.069*
	Pre-test Length	2	0.924
	Residual	50	
Discrimination **y_c**	Outcome Condition	4	0.302
	Sample Size	3	38.271*
	Pre-test Length	2	3.917
	Residual	50	
Discrimination **y_t**	Outcome Condition	4	0.235
	Sample Size	3	58.024*
	Pre-test Length	2	0.931
	Residual	50	
Difficulty **$(y_t$-$y_c)$**	Outcome Condition	4	8.426*
	Sample Size	3	4.565*
	Pre-test Length	2	1.263
	Residual	50	

Note: * $p < 0.05$

Table 8

ANCOVA table for the RMSE in the ability-based scenario.

Response	**Source**	**DF**	**F**
Difficulty y_c	Difficulty y_c	1	151.360*
	Outcome Condition	4	155.231*
	Sample Size	3	103.494*
	Pre-test Length	2	0.979
	Residual	49	
Difficulty y_t	Difficulty y_t	1	191.870*
	Outcome Condition	4	161.911*
	Sample Size	3	126.784*
	Pre-test Length	2	4.252*
	Residual	49	
Discrimination y_c	Discrimination y_c	1	2256.892*
	Outcome Condition	4	2.854*
	Sample Size	3	195.319*
	Pre-test Length	2	51.777*
	Residual	49	
Discrimination y_t	Discrimination y_t	1	2232.399*
	Outcome Condition	4	2.195
	Sample Size	3	133.738*
	Pre-test Length	2	51.839*
	Residual	49	
Difficulty (y_t-y_c)	Difficulty difference (y_t-y_c)	1	207.695*
	Outcome Condition	4	192.142*
	Sample Size	3	119.590*
	Pre-test Length	2	2.430
	Residual	49	

Note: * $p < 0.05$

Quantifying the Treatment Difference The final part of the comparison under the ability-based scenario was to quantify the treatment effect found by the IRT method and compare these treatment estimates to the estimates found by the other methods. Since the outcome in each group is treated as an item in the overall IRT model, item characteristic curves (ICC) can be created using the item parameters. These two curves are plotted over the entire ability spectrum. For an illustration of what this looks like, see Figure 9. The area between these curves represents the difference between the treatment and control groups. This difference is the treatment effect. To quantify this treatment effect, integration can be used to find the area. Furthermore, including the decision variable in the IRT model, which is done under the ability-based scenario, assumes that there is a distribution difference between the group assignment, and this integration must be weighted by the proportion of people in the treatment group.

Mathematically the difference ($\Delta(u)$) between the two curves can be found by using the following formula. Here, a_{yc} represents the estimated discrimination of the outcome for the control group, a_{yt} represents the estimated discrimination of the outcome for the treatment group, b_{yc} represents the estimated difficulty of the outcome for the control group, b_{yt} represents the estimated difficulty of the outcome for the treatment group, and u represents a point on the ability scale.

$$\Delta(u) = \frac{-1}{1+e^{-a_{yc}(u-b_{yc})}} + \frac{-1}{1+e^{-a_{yt}(u-b_{yt})}}$$

This difference is then weighted by the distribution of the decision variable and the proportion assigned to treatment, $wt\ (u)$.

$$wt(u) = \frac{1 \Big/ \left(\sqrt{2\pi}\right) e^{-\frac{1}{2}u^2}}{1+e^{-a_D(u-b_D)} prop(treatment)}$$

Finally, the weighted difference is integrated, over the conditional distribution of the population of interest, *S*, from negative infinity to positive infinity to quantify the treatment effect, *eff* (*S*).

$$eff(S) = \int_{-\infty}^{\infty} \Delta(u) wt(u) f(u \mid S) du$$

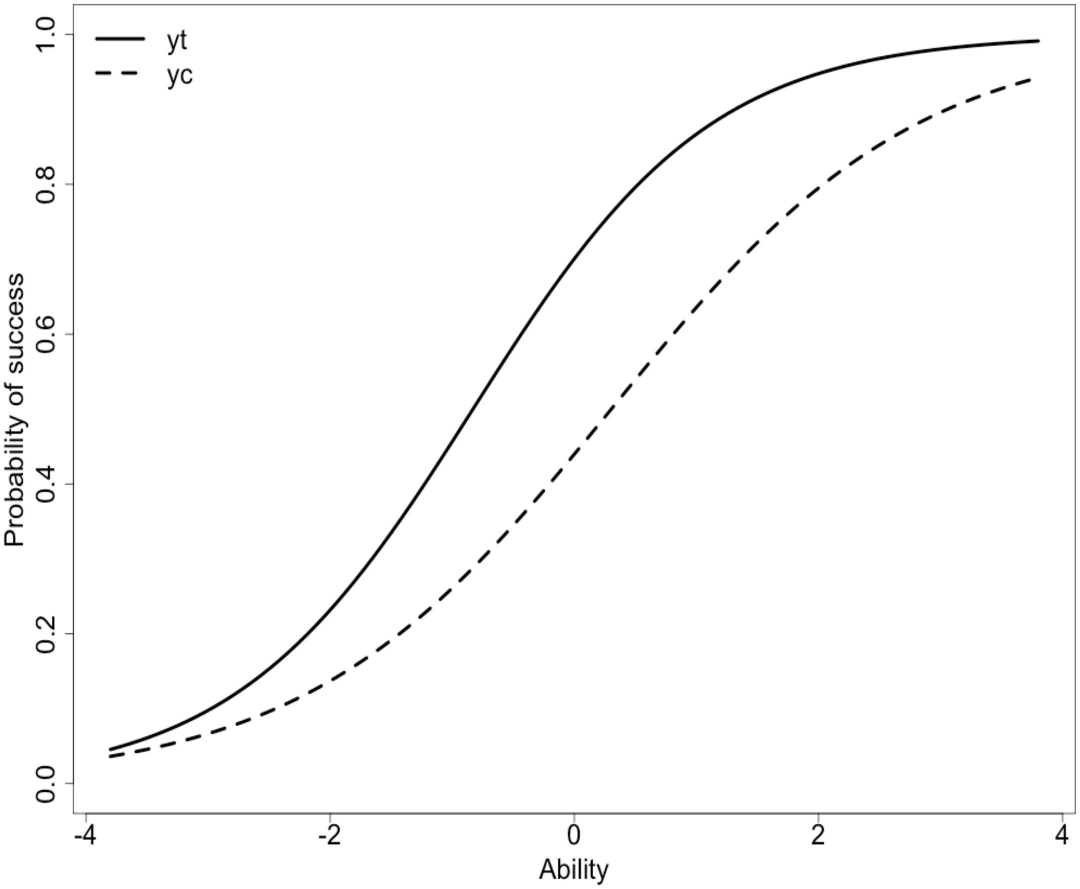

Figure 9: Hypothetical example of ICC outcome curves.

For each of the three methods, this treatment effect was calculated for each simulation condition. For the No Difference group, the effect should be zero since both the treatment and control group found the outcome equally difficult. Both negative outcome conditions (Low - and

High -) should be negative, indicating that there is a negative treatment effect because the control group finds the outcome easier than the treatment group. This value should be larger in magnitude for the High - outcome condition since a greater difference in difficulty exists initially between groups. The positive outcome conditions (Low + and High +) should both show positive effects since the treatment group found the outcome measure easier than the control group. This value should be larger in magnitude for the High + outcome condition since there was a greater difference in difficulty between groups. Table 9 through Table 12 show the means and standard deviations by method for the treatment effect.

The true values were calculated for each condition. Since the difficulty value of the ability-based treatment decision is zero and the discrimination is one, the proportion expected in treatment is 50%. Using the initial values for the outcome conditions, the true treatment effect for each outcome condition is as follows—No Difference: treatment effect = 0; Low -: treatment effect = -0.204; Low +: treatment effect = 0.204; High -: treatment effect = -0.393; and High +: treatment effect = 0.393. For the IRT method, these values follow as predicted shown by mean values very close to the true values and small standard deviations. For the LR and propensity score methods, the direction of effects follow as expected, but the magnitude does not. Both the estimates from LR and propensity score are further from the true values than the estimates from the IRT method. In addition, both methods show more variation, shown by larger standard deviations, than the IRT method.

Table 9

Means (and standard deviations) of treatment effect estimates, by method, for N = 500.

X	OC	IRT	LR	PS
10	**No Difference**	0.001 (0.056)	0.319 (0.219)	0.074 (0.065)
	Low -	-0.203 (0.049)	-0.623 (0.213)	-0.136 (0.059)
	Low +	0.203 (0.059)	1.265 (0.230)	0.275 (0.065)
	High -	-0.395 (0.048)	-1.584 (0.254)	-0.339 (0.057)
	High +	0.393 (0.058)	2.214 (0.246)	0.460 (0.064)
20	**No Difference**	-0.001 (0.051)	0.142 (0.241)	0.038 (0.068)
	Low -	-0.203 (0.048)	-0.836 (0.251)	-0.166 (0.062)
	Low +	0.204 (0.052)	1.129 (0.239)	0.236 (0.067)
	High -	-0.392 (0.042)	-1.823 (0.272)	-0.361 (0.056)
	High +	0.395 (0.051)	2.131 (0.252)	0.427 (0.067)
50	**No Difference**	-0.001 (0.047)	0.031 (0.244)	0.014 (0.075)
	Low -	-0.205 (0.046)	-0.986 (0.269)	-0.189 (0.071)
	Low +	0.202 (0.048)	1.042 (0.239)	0.217 (0.081)
	High -	-0.394 (0.040)	-2.013 (0.298)	-0.382 (0.063)
	High +	0.392 (0.050)	2.061 (0.263)	0.404 (0.076)

Note: True values for the outcome conditions are No Difference = 0.000, Low - = -0.204, Low + = 0.204, High - = -0.393, High + = 0.393.

Table 10

Means (and standard deviations) of treatment effect estimates, by method, for N=1,000.

X	OC	IRT	LR	PS
10	**No Difference**	0.002 (0.039)	0.325 (0.151)	0.073 (0.043)
	Low -	-0.205 (0.036)	-0.625 (0.156)	-0.139 (0.040)
	Low +	0.203 (0.041)	1.257 (0.157)	0.274 (0.041)
	High -	-0.393 (0.034)	-1.563 (0.175)	-0.336 (0.038)
	High +	0.395 (0.039)	2.208 (0.164)	0.457 (0.043)
20	**No Difference**	0.000 (0.034)	0.151 (0.158)	0.038 (0.045)
	Low -	-0.205 (0.034)	-0.839 (0.178)	-0.171 (0.041)
	Low +	0.205 (0.036)	1.134 (0.163)	0.241 (0.045)
	High -	-0.392 (0.030)	-1.812 (0.192)	-0.362 (0.038)
	High +	0.393 (0.037)	2.112 (0.182)	0.428 (0.047)
50	**No Difference**	-0.002 (0.035)	0.032 (0.177)	0.014 (0.049)
	Low -	-0.203 (0.032)	-0.966 (0.182)	-0.186 (0.043)
	Low +	0.203 (0.035)	1.043 (0.175)	0.218 (0.050)
	High -	-0.393 (0.029)	-1.990 (0.207)	-0.381 (0.040)
	High +	0.395 (0.033)	2.063 (0.179)	0.412 (0.047)

Note: True values for the outcome conditions are No Difference = 0.000, Low - = -0.204, Low + = 0.204, High - = -0.393, High + = 0.393.

Table 11

Means (and standard deviations) of treatment effect estimates, by method, for N=5,000.

X	OC	IRT	LR	PS
10	**No Difference**	0.000 (0.017)	0.317 (0.068)	0.070 (0.018)
	Low -	-0.205 (0.016)	-0.620 (0.071)	-0.140 (0.017)
	Low +	0.204 (0.018)	1.257 (0.069)	0.274 (0.018)
	High -	-0.393 (0.015)	-1.551 (0.075)	-0.336 (0.015)
	High +	0.394 (0.018)	2.197 (0.074)	0.457 (0.017)
20	**No Difference**	-0.001 (0.016)	0.146 (0.074)	0.035 (0.019)
	Low -	-0.205 (0.014)	-0.832 (0.076)	-0.171 (0.017)
	Low +	0.204 (0.016)	1.125 (0.075)	0.240 (0.018)
	High -	-0.394 (0.014)	-1.811 (0.090)	-0.363 (0.016)
	High +	0.394 (0.016)	2.104 (0.081)	0.426 (0.019)
50	**No Difference**	0.000 (0.015)	0.045 (0.074)	0.016 (0.018)
	Low -	-0.204 (0.014)	-0.961 (0.079)	-0.189 (0.016)
	Low +	0.204 (0.016)	1.046 (0.077)	0.219 (0.019)
	High -	-0.393 (0.013)	-1.971 (0.090)	-0.381 (0.015)
	High +	0.393 (0.015)	2.046 (0.081)	0.408 (0.019)

Note: True values for the outcome conditions are No Difference = 0.000, Low - = -0.204, Low + = 0.204, High - = -0.393, High + = 0.393.

Table 12

Means (and standard deviations) of treatment effect estimates, by method, for N=10,000.

X	OC	IRT	LR	PS
10	**No Difference**	0.000 (0.012)	0.317 (0.047)	0.070 (0.012)
	Low -	-0.204 (0.012)	-0.614 (0.050)	-0.138 (0.011)
	Low +	0.204 (0.013)	1.253 (0.049)	0.273 (0.012)
	High -	-0.394 (0.011)	-1.555 (0.054)	-0.337 (0.010)
	High +	0.393 (0.013)	2.193 (0.053)	0.457 (0.011)
20	**No Difference**	0.000 (0.011)	0.148 (0.054)	0.036 (0.012)
	Low -	-0.205 (0.011)	-0.830 (0.056)	-0.170 (0.012)
	Low +	0.204 (0.012)	1.123 (0.053)	0.240 (0.012)
	High -	-0.394 (0.011)	-1.809 (0.061)	-0.364 (0.010)
	High +	0.393 (0.013)	2.099 (0.058)	0.426 (0.013)
50	**No Difference**	0.000 (0.011)	0.045 (0.054)	0.016 (0.013)
	Low -	-0.204 (0.010)	-0.962 (0.058)	-0.189 (0.012)
	Low +	0.204 (0.011)	1.044 (0.053)	0.219 (0.012)
	High -	-0.394 (0.010)	-1.975 (0.068)	-0.381 (0.011)
	High +	0.394 (0.012)	2.047 (0.056)	0.409 (0.013)

Note: True values for the outcome conditions are No Difference = 0.000, Low - = -0.204, Low + = 0.204, High - = -0.393, High + = 0.393.

To visualize the difference between the estimated effect and the true value, bias and RMSE calculations were performed. Figure 10 through Figure 12 show plots of the bias of these treatment effects and Figure 13 through Figure 15 show plots of the RMSE. For comparison, within bias and RMSE, these figures are plotted on the same scale. However, it is important to note that the range of values varied by method. For the IRT method, bias ranged from -0.002 to 0.002 and RMSE from 0.000 to 0.060. For LR, bias ranged from -1.620 to 1.821 and RMSE from 0.000 to 1.840. Finally for propensity score matching, bias ranged from 0.01 to 0.08 and RMSE from 0.000 to 0.100.

From Figure 10 and Figure 13 it can be seen that the bias and RMSE of the estimates from the IRT method are very small, indicating that this method returns accurate values of the treatment effect. There seems to be a slight improvement in estimation as sample size increases, but since the values are so close to zero it is hard to describe. From Figure 11 and Figure 14 it can be seen that the LR method has difficulty estimating the true treatment effect. This difficulty is seen by the high values of bias and RMSE. From the plot of the bias, LR struggles in all treatment conditions, but especially High - and High +. There does not seem to be an improvement in estimation as sample size gets larger and it seems as if better estimates are produced when the pre-test is shortest. From Figure 12 and Figure 15 it can be seen that the propensity score method estimates the treatment effect almost as well as the IRT method. Yet there is deviation from zero shown in both the bias and RMSE plots. It also seems as if as sample size increases, the estimates from the longer tests get better while the estimates from the shorter tests remain the same.

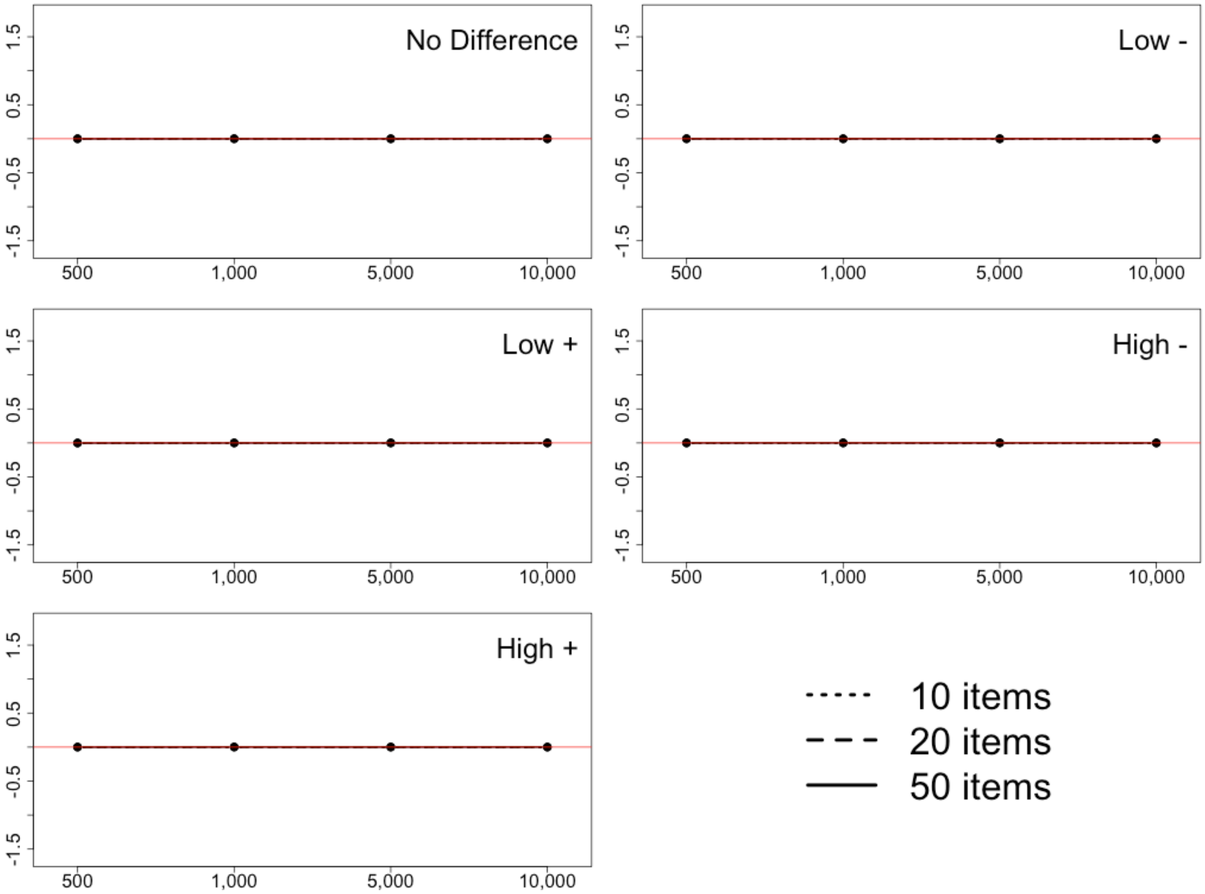

Figure 10: Profile plot of bias by outcome condition for the IRT method under ability-based scenario.

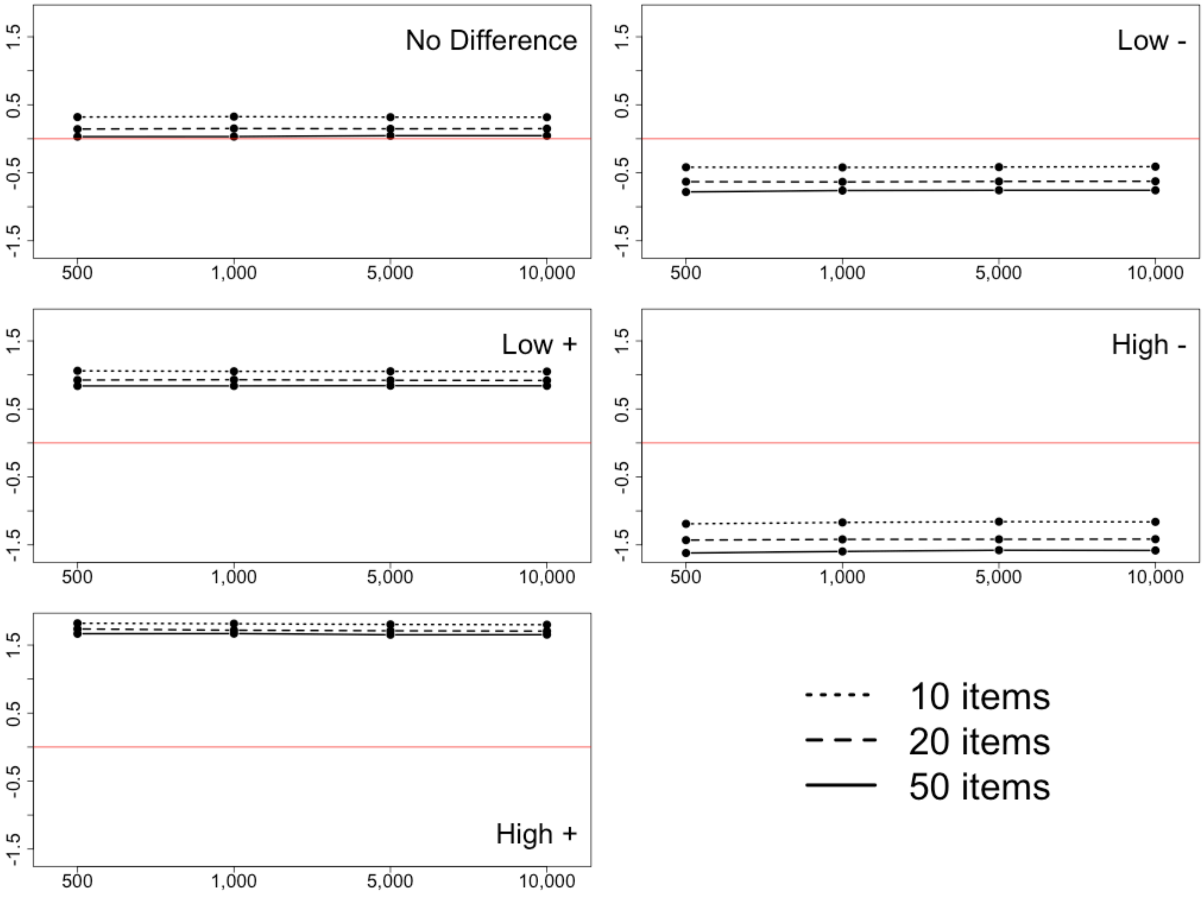

Figure 11: Profile plot of bias by outcome condition for the LR method under ability-based scenario.

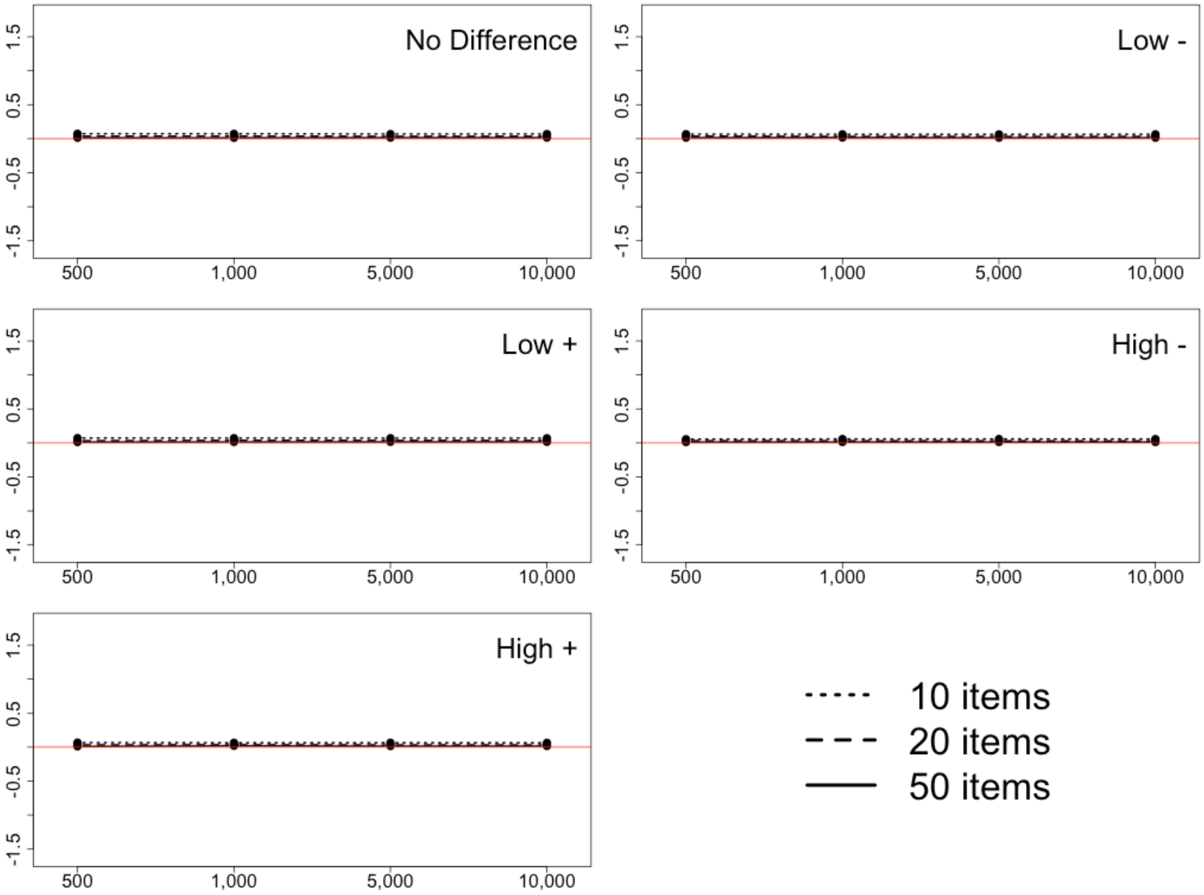

Figure 12: Profile plot of bias by outcome condition for the propensity score method under ability-based scenario.

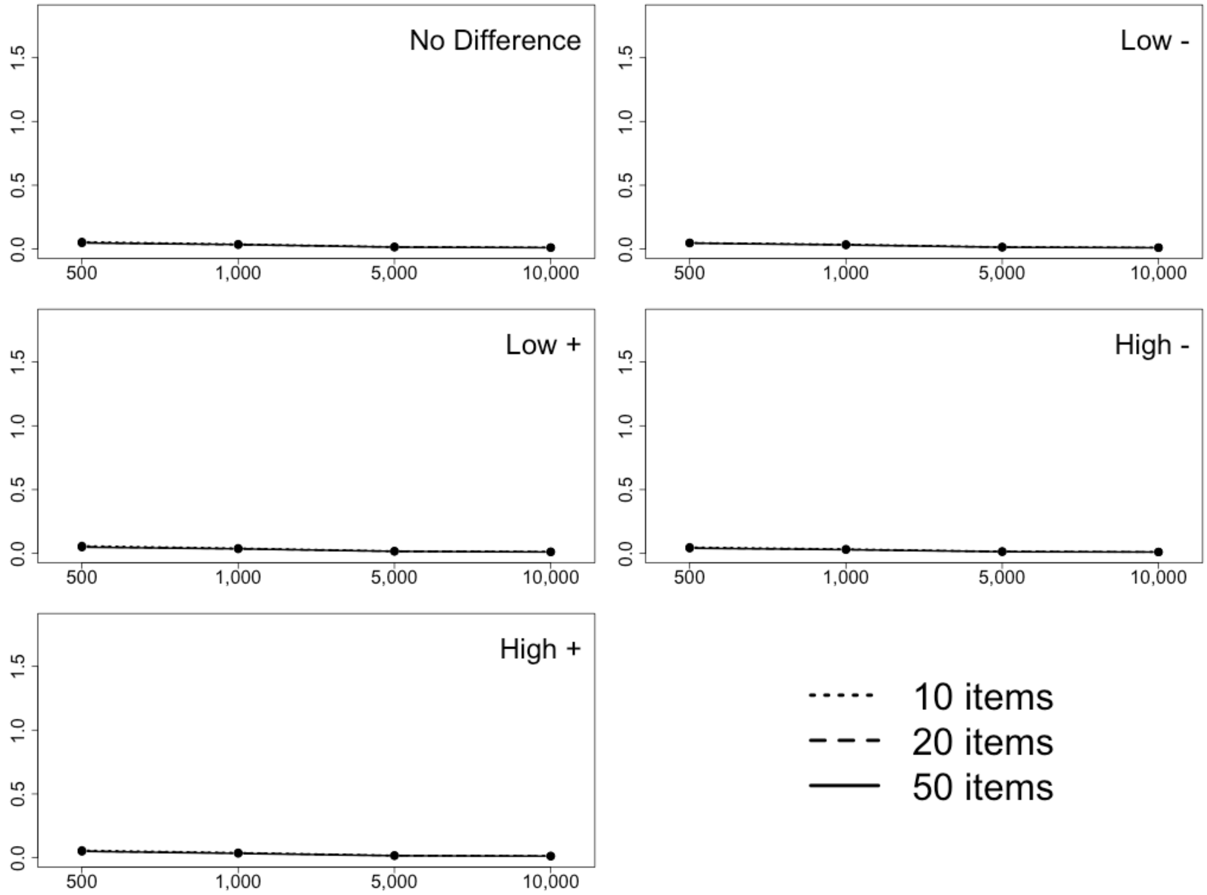

Figure 13: Profile plot of RMSE by outcome condition for the IRT method, under the ability-based scenario.

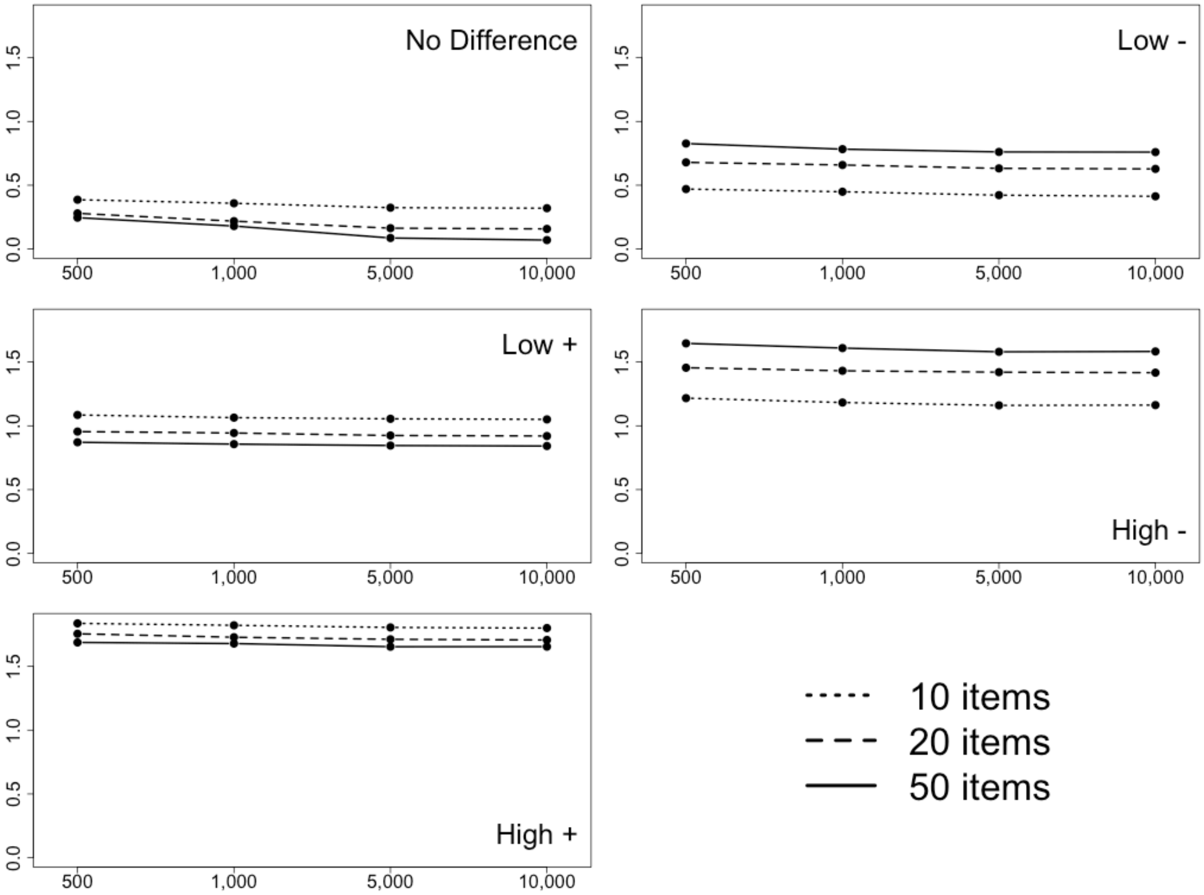

Figure 14: Profile plot of RMSE by outcome condition for the LR method, under the ability-based scenario

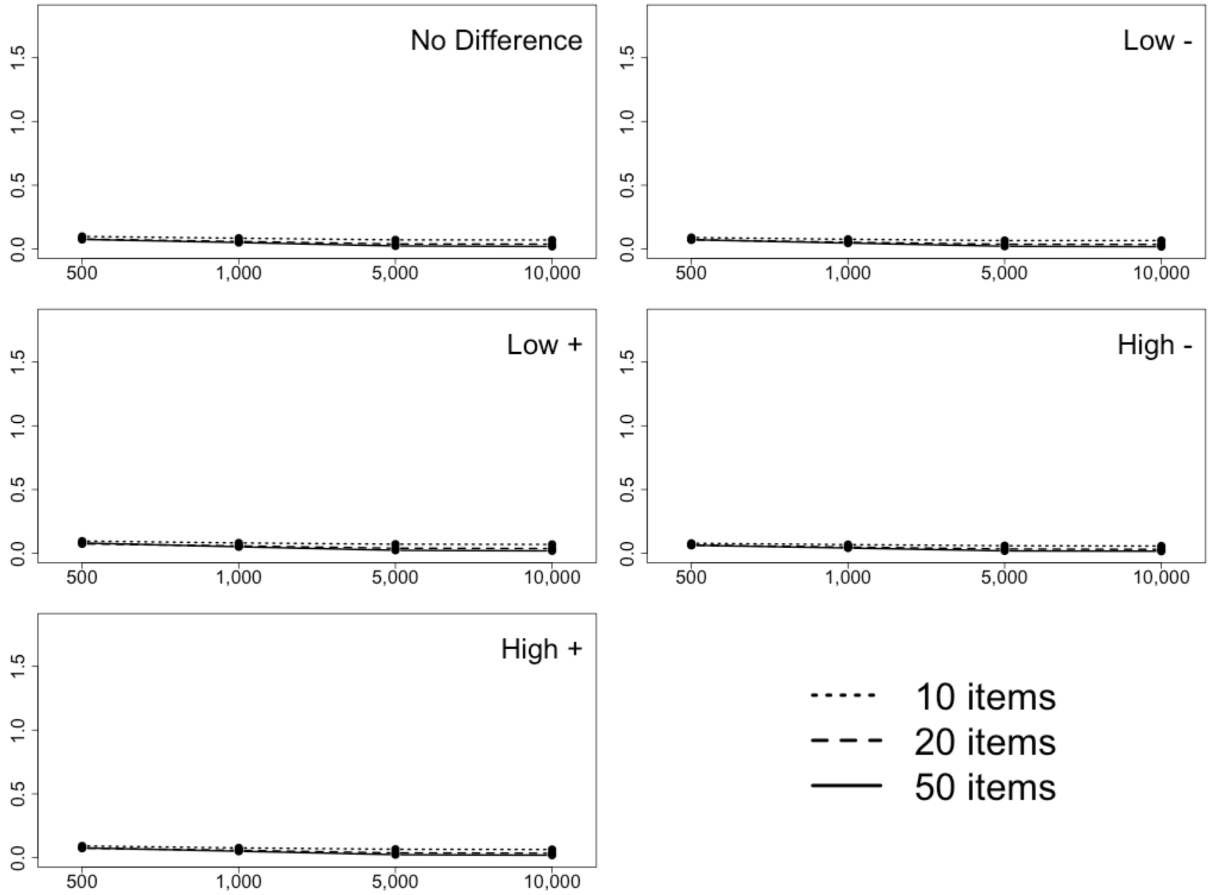

Figure 15: Profile plot of RMSE by outcome condition for the propensity score method, under the ability-based scenario

Comparison Under the Cut-score Scenario

Recall that the cut-score scenario is the model where the decision is based strictly on the pre-test (see Figure 2). In this scenario, students are placed into treatment based on a cut-score of 50% accuracy on the test. The mean parameter estimates for the outcome variables in the cut-score scenario, along with the true values, RMSE, and bias can be seen in Table A6 through Table A9 of the Appendix. The estimates for the control group produced by this model do not look reasonable. There are some very extreme values that could indicate that the model does not fit this data very well. These patterns can be further elaborated visually.

Type I and Type II Error Rates Recall that the main hypothesis of this study is to determine how the IRT based method compares to existing methods of causal inference. For the three models—LR, RD, and IRT—the proportion of simulated datasets where statistically significant effects, $p \leq 0.05$, were found was calculated for each simulation condition. Table 13 through Table 16 show the proportion of significant differences found by each method for the outcome conditions.

The No Difference outcome condition represents the situation where both groups find the outcome equally difficult and no true differences exist, indicating that differences between the data should not be found more than by chance. Here, we would expect to see a proportion of 0.05 if the causal inference method was performing as expected. This condition represents the Type I error rate. Across all four sample sizes the IRT method has good Type I error rates as shown by a lower proportion of significant differences found by this model. LR and RD have Type I error rates that are higher than those of the IRT model.

Recall that the conditions where true values exist represent the power of the test and that a good test has power near one showing its ability at detecting true differences (Casella & Berger, 2002, p. 383). LR and RD show good power across the different sample sizes. However, the IRT method has poor power in this scenario, which can be attributed to the previously discussed fitting issues.

Table 13

Proportion of statistically significant effects found by each model, under the cut-score scenario, for a sample size of 500.

Simulation Characteristics		**IRT**	**LR**		**RD**	
X	**OC**		**Decision**	**Interaction**	**Decision**	**Interaction**
10	**No Difference**	0.063	0.041	0.045	0.043	0.045
	Low -	0.088	0.143	0.039	0.331	0.039
	Low +	0.055	0.152	0.042	0.493	0.042
	High -	0.123	0.287	0.041	0.749	0.041
	High +	0.028	0.617	0.046	0.978	0.046
20	**No Difference**	0.042	0.072	0.101	0.053	0.052
	Low -	0.074	0.151	0.069	0.352	0.055
	Low +	0.037	0.507	0.502	0.697	0.062
	High -	0.109	0.644	0.273	0.819	0.041
	High +	0.015	0.911	0.907	0.997	0.057
50	**No Difference**	0.015	0.064	0.017	0.073	0.050
	Low -	0.025	0.028	0.015	0.397	0.050
	Low +	0.010	0.395	0.039	0.845	0.069
	High -	0.063	0.100	0.008	0.949	0.055
	High +	0.003	0.794	0.777	0.999	0.086

Note: X represents the test length and OC represents the outcome condition.

Table 14

Proportion of statistically significant effects found by each model, under the cut-score scenario, for a sample size of 1,000.

Simulation Characteristics		**IRT**	**LR**		**RD**	
X	**OC**		**Decision**	**Interaction**	**Decision**	**Interaction**
10	**No Difference**	0.025	0.048	0.055	0.050	0.055
	Low -	0.038	0.255	0.055	0.611	0.055
	Low +	0.014	0.268	0.059	0.793	0.059
	High -	0.055	0.493	0.046	0.950	0.046
	High +	0.008	0.874	0.061	1.000	0.061
20	**No Difference**	0.016	0.117	0.168	0.066	0.066
	Low -	0.016	0.321	0.127	0.599	0.042
	Low +	0.004	0.810	0.790	0.926	0.071
	High -	0.045	0.901	0.487	0.978	0.069
	High +	0.003	0.997	0.995	1.000	0.081
50	**No Difference**	0.003	0.182	0.043	0.091	0.055
	Low -	0.007	0.037	0.023	0.756	0.063
	Low +	0.119	0.736	0.071	0.992	0.069
	High -	0.017	0.148	0.023	1.000	0.054
	High +	1.000	0.986	0.148	1.000	0.099

Note: X represents the test length and OC represents the outcome condition.

Table 15

Proportion of statistically significant effects found by each model, under the cut-score scenario, for a sample size of 5,000.

Simulation Characteristics		**IRT**	**LR**		**RD**	
X	**OC**		**Decision**	**Interaction**	**Decision**	**Interaction**
10	**No Difference**	0.047	0.047	0.059	0.053	0.059
	Low -	0.588	0.757	0.042	0.999	0.042
	Low +	0.998	0.911	0.073	1.000	0.073
	High -	0.790	0.984	0.059	1.000	0.059
	High +	1.000	1.000	0.088	1.000	0.088
20	**No Difference**	0.051	0.323	0.580	0.133	0.126
	Low -	0.444	0.929	0.311	0.998	0.070
	Low +	0.998	1.000	1.000	1.000	0.127
	High -	0.119	1.000	0.983	1.000	0.065
	High +	1.000	1.000	1.000	1.000	0.202
50	**No Difference**	0.049	0.715	0.120	0.207	0.162
	Low -	0.844	0.051	0.056	1.000	0.113
	Low +	1.000	0.999	0.310	1.000	0.195
	High -	0.990	0.538	0.034	1.000	0.085
	High +	1.000	1.000	0.568	1.000	0.269

Note: X represents the test length and OC represents the outcome condition.

Table 16

Proportion of statistically significant effects found by each model, under the cut-score scenario, for a sample size of 10,000.

Simulation Characteristics		**IRT**	**LR**		**RD**	
X	**OC**		**Decision**	**Interaction**	**Decision**	**Interaction**
10	**No Difference**	0.057	0.044	0.058	0.044	0.058
	Low -	0.978	0.973	0.056	1.000	0.056
	Low +	1.000	0.997	0.076	1.000	0.076
	High -	1.000	1.000	0.043	1.000	0.043
	High +	1.000	1.000	0.104	1.000	0.104
20	**No Difference**	0.055	0.561	0.899	0.187	0.178
	Low -	0.937	0.994	0.511	1.000	0.129
	Low +	1.000	1.000	1.000	1.000	0.298
	High -	1.000	1.000	1.000	1.000	0.080
	High +	1.000	1.000	1.000	1.000	0.391
50	**No Difference**	0.054	0.951	0.255	0.438	0.255
	Low -	1.000	0.058	0.086	1.000	0.166
	Low +	1.000	1.000	0.537	1.000	0.359
	High -	1.000	0.839	0.040	1.000	0.098
	High +	1.000	1.000	0.865	1.000	0.450

Note: X represents the test length and OC represents the outcome condition.

To visualize this further, the proportion of statistically significant effects from each condition of the three methods were plotted in a histogram by each treatment effect outcome group. Figure 16 shows this plot for the No Difference condition under the IRT method. As stated previously, *p*-values that fall under the null hypothesis are uniformly distributed (Casella and Berger, 2002, p. 397-8). Since this outcome condition represents the situation where both treatment and control groups find the outcome equally difficult, which is the null hypothesis, these histograms should look like a uniform distribution. As alluded to in Table 13 through Table 16, this figure shows that model fitting problems existed especially for the two smaller sample sizes of 500 and 1,000. The four other plots for the remaining outcome conditions for the IRT model can be seen in the Appendix in Figure A31 through Figure A34. Figure 17 shows these same No Difference outcome condition plots for the LR model. Figure 18 shows the No Difference outcome plots for the RD model.

Within each of these figures, there are some simulation condition combinations of sample size and test length where the distribution of the proportion of significant effects look uniform. However, more frequently it appears that these histograms do not look uniform. This helps to visualize the amount of Type I error that exists in these existing causal inference methods when applied to data of this type. For both the LR method and the RD shorter tests show more uniform distributions. This could indicate that the Type I error is more biased with measurement error in these conditions. The remaining plots for the other outcome conditions can be seen in the Appendix--Figure A35 through Figure A38 for the LR method and Figure A39 through Figure A42 for the RD method.

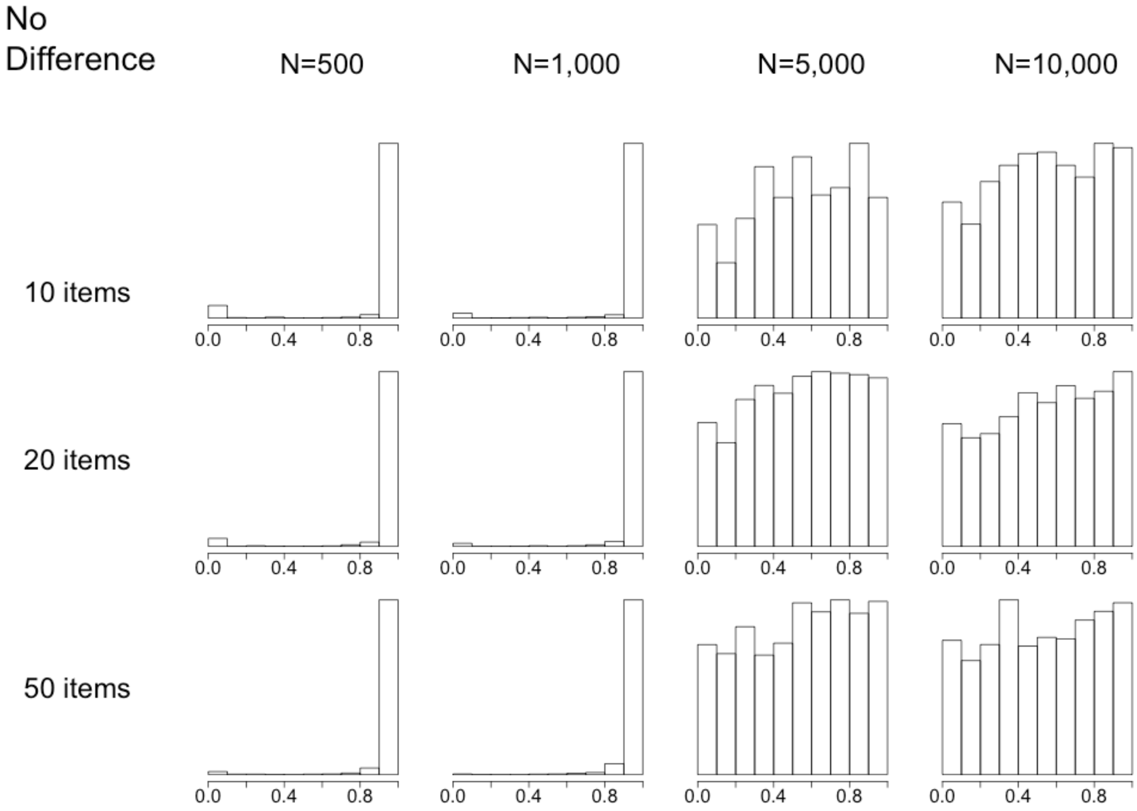

Figure 16: Histograms of the *p*-values from the IRT method in the cut-score scenario for the No Difference outcome condition.

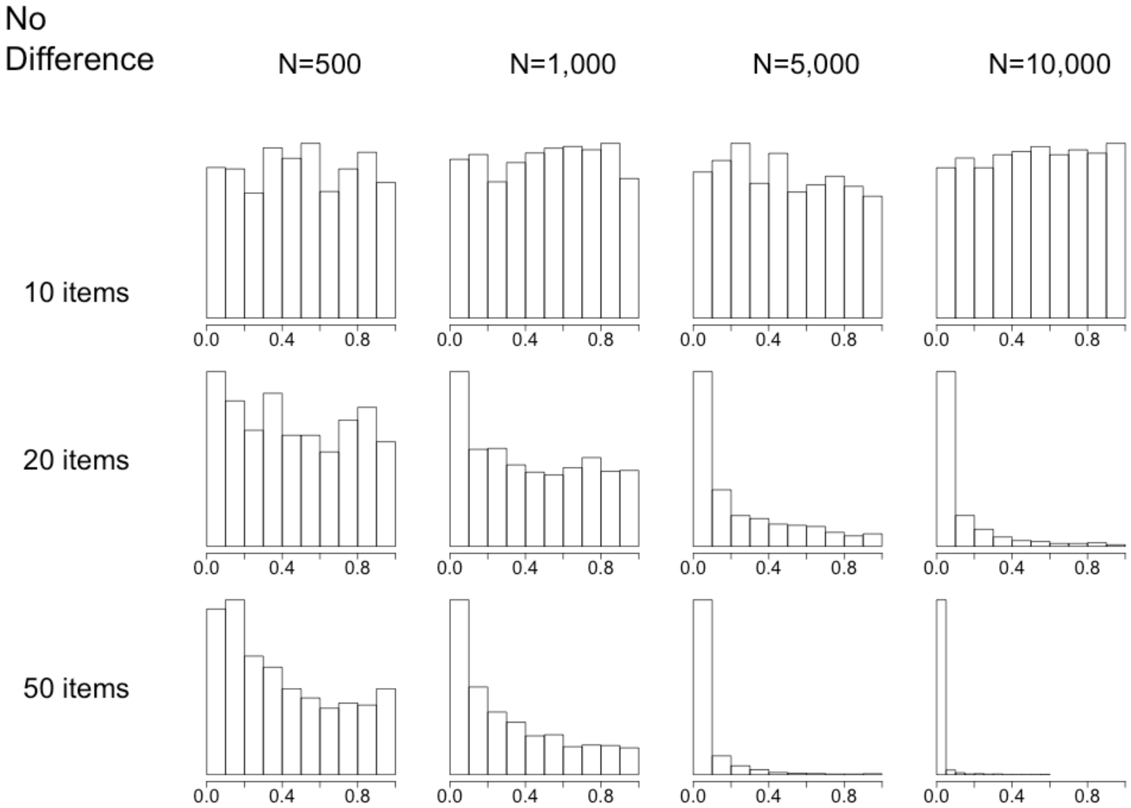

Figure 17: Histograms of the *p*-values from the LR method under the cut-score scenario for the No Difference outcome condition.

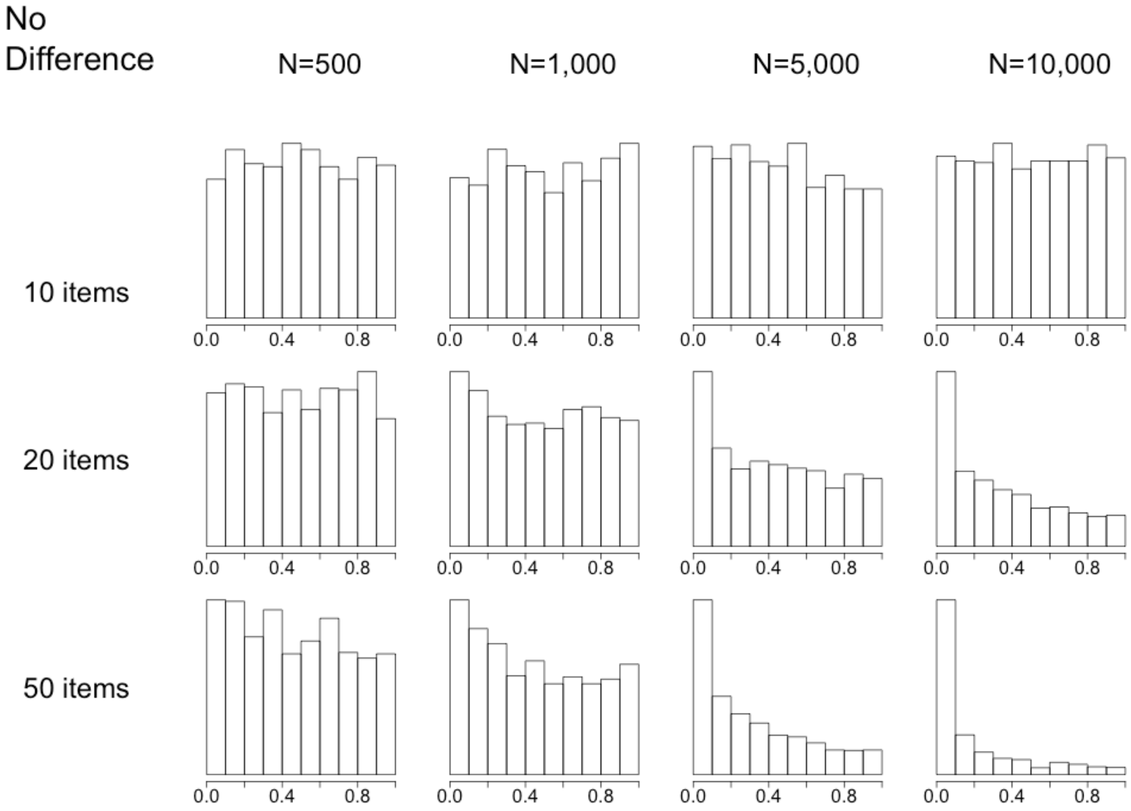

Figure 18: Histograms of the *p*-values from the RD method under the cut-score scenario for the No Difference outcome condition.

Bias and RMSE for the IRT Method As in the other scenario, to evaluate the IRT method in terms of parameter recovery, biases and RMSE values were calculated for each of the parameters of the outcome. Then, these values were plotted, seen in Figure 19 and Figure 20. As discussed throughout this dissertation, this model encountered problems while being fit to the data. At times, extreme parameter estimates were produced by the IRT model. This is reflected in the following figures.

Figure 19 shows the bias for the difficulty parameter of the outcome variable in the treatment group. Overall, the parameter recovery for this specific parameter does not seem bad.

Across these plots emerges a trend—better parameter recovery is seen as the sample size increases. The largest change in bias can be seen in the panel representing the High - outcome condition. Recall that this condition is where the treatment group found the outcome to be very difficult. Also recall that these students were placed into treatment if their total test score fell below the cut score value corresponding to half the items on the test. Therefore, together this means that these students are struggling and are given an outcome that is very difficult for them. The control group shows the same thing for the High - outcome condition for the difficulty (Figure A43 in the Appendix). Even the discrimination parameter, which was not manipulated, shows this trend for the High - outcome condition (Figure A44 for the treatment and Figure A45 for the control group in the Appendix). This could explain why the parameter recovery is not very good in this condition.

Figure 20 shows the RMSE for the difficulty parameter of the outcome in the treatment group. The panel that shows the largest change in bias is once again the panel corresponding to the High - outcome condition. The RMSE seem acceptable and do not exceed 0.5. However, the RMSE for the control group is poor for all conditions, especially in the High - outcome condition (see Figure A46 in the Appendix). The RMSEs for the discrimination in both treatment and control groups reduce as sample size increases, but are poor (i.e. large) for the High - outcome condition (Figure A47 for the treatment group and Figure A48 for the control group in the Appendix). Also, for completeness, scatterplots of the bias and RMSE for all of the remaining items can be seen in the Appendix. See Figure A49 to Figure A55 for bias and Figure A56 to Figure A61 for RMSE.

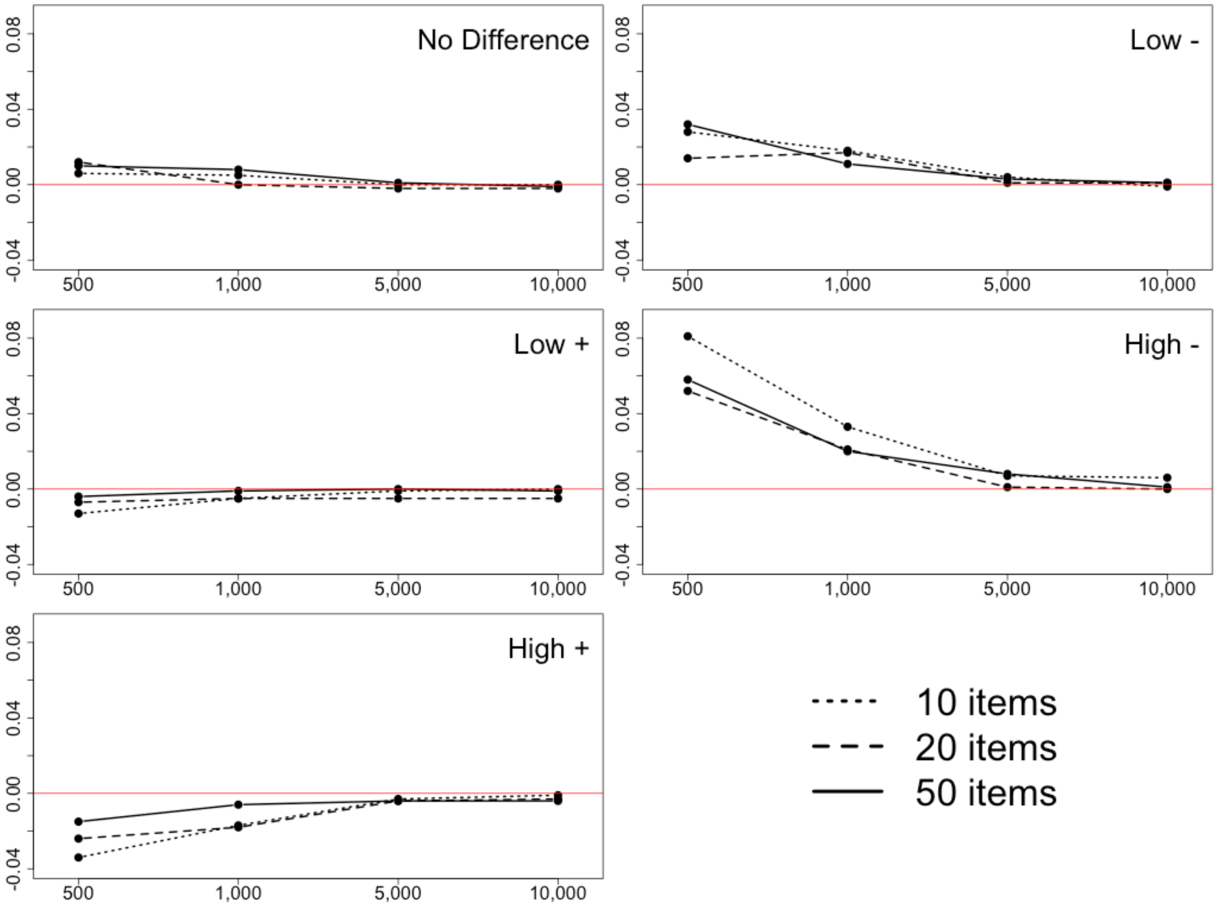

Figure 19: Profile plot for the bias of the difficulty parameter of the outcome variable for the treatment group under the cut-score scenario.

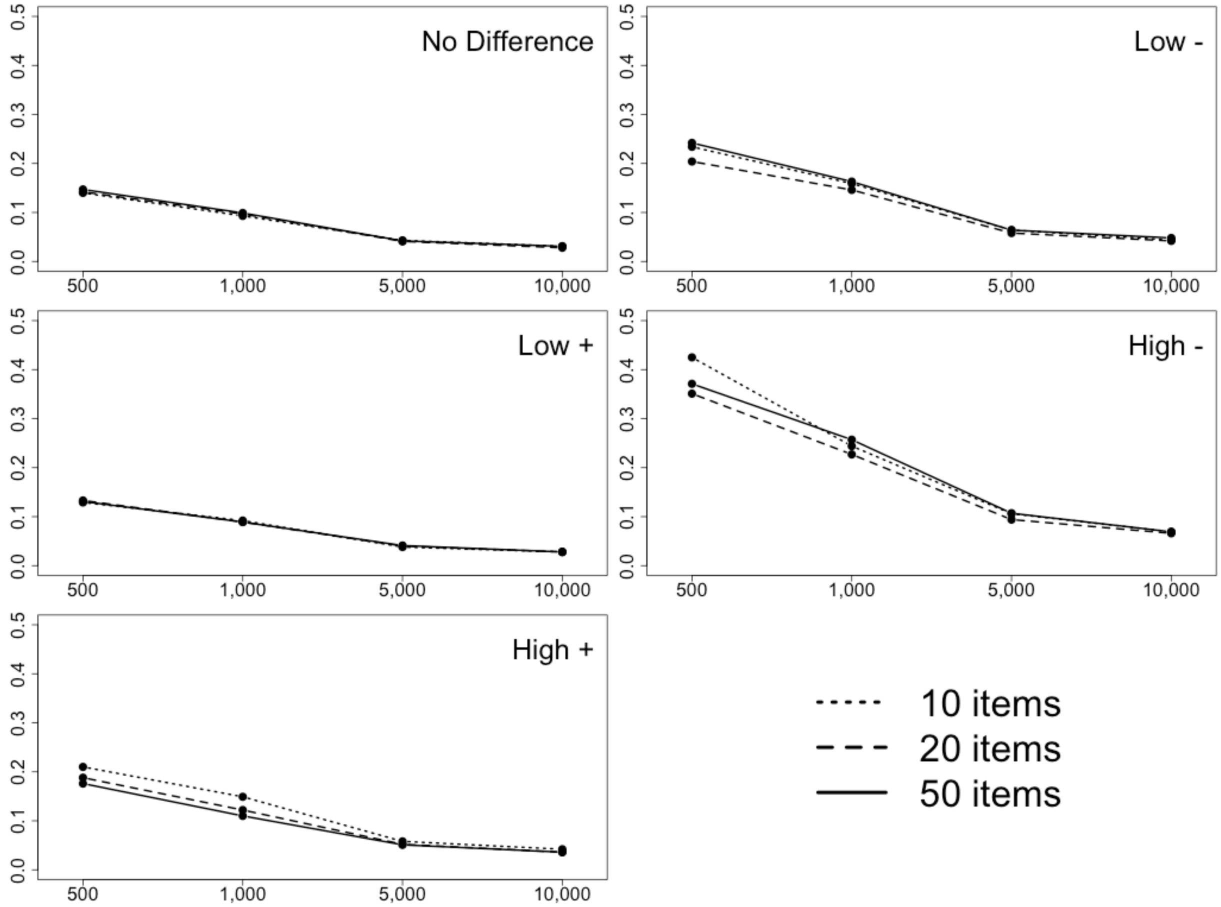

Figure 20: Profile plot for the RMSE of the difficulty parameter of the outcome variable for the treatment group under the cut-score scenario.

ANOVA and ANCOVA Analysis of Simulation Conditions The previous profile plots tell a similar story to those under the ability-based scenario. They indicate that there could be an under lying association between aspects of the simulation conditions and the bias and RMSE values. Once again, ANOVA and ANCOVA analyses were performed to gain insight into which aspects of the simulations were significantly related to the IRT method's ability to recover parameters.

Again, individual three-way factorial design ANCOVA models were run to see if certain components of each condition—sample size, pre-test length, outcome condition (No Difference, Low -, Low +, High -, High +)—explained the variation in bias and RMSE values while controlling for parameter estimates. These three-way factorial models had equal groups in each cell of the design, making effect interpretation easier. Five models used a corresponding bias as an outcome and fived used a corresponding RMSE.

ANCOVA models with full interactions were estimated, but had "an essentially perfect fit" according to R and were unreliable. Therefore, the interactions were removed and only main effect models were fit. For the bias, these main effect ANCOVA models had the same "essentially perfect fit" and therefore the parameter estimate portion of the model (difficulty, discrimination, or difference in difficult) was removed from each model and an ANOVA model was fit instead. The results of these ANOVA and ANCOVA analyses are shown in Table 17 Table 18.

Table 17 shows the association with bias. Sample size is always significantly associated with the bias, regardless of the parameter being estimated. The outcome condition is significantly associated with bias only for the difficulty parameter. Pre-test length was significantly associated with the bias of the difficulty parameter for the treatment group and with the difference of the difficulty between the treatment and control groups. Table 18 shows the results of the RMSE analysis. Once again, sample size is always a significant factor. Outcome condition is significantly related in all but the RMSE of the discrimination parameter of the control group; yet, pre-test length is rarely significantly associated.

Table 17

ANOVA table for the bias in the cut-score scenario.

Response	Source	DF	F
Difficulty y_c	Outcome Condition	4	6.464*
	Sample Size	3	20.303*
	Pre-test Length	2	3.272*
	Residual	50	
Difficulty y_t	Outcome Condition	4	12.427*
	Sample Size	3	3.297*
	Pre-test Length	2	0.475
	Residual	50	
Discrimination y_c	Outcome Condition	4	2.032
	Sample Size	3	4.123*
	Pre-test Length	2	2.135
	Residual	50	
Discrimination y_t	Outcome Condition	4	2.175
	Sample Size	3	38.475*
	Pre-test Length	2	2.149
	Residual	50	
Difficulty (y_t-y_c)	Outcome Condition	4	6.484*
	Sample Size	3	20.257*
	Pre-test Length	2	3.255*
	Residual	50	

Note: * $p < 0.05$

Table 18

ANCOVA table for the RMSE in the cut-score scenario.

Response	Source	DF	F
Difficulty y_c	Difficulty y_c	1	1880.129*
	Outcome Condition	4	4.818*
	Sample Size	3	32.697*
	Pre-test Length	2	1.715
	Residual	49	
Difficulty y_t	Difficulty y_t	1	155.425*
	Outcome Condition	4	104.334*
	Sample Size	3	119.376*
	Pre-test Length	2	0.300
	Residual	49	
Discrimination y_c	Discrimination y_c	1	1450.687*
	Outcome Condition	4	0.669
	Sample Size	3	2.856*
	Pre-test Length	2	2.956
	Residual	49	
Discrimination y_t	Discrimination y_t	1	2647.723*
	Outcome Condition	4	7.148*
	Sample Size	3	308.138*
	Pre-test Length	2	48.804*
	Residual	49	
Difficulty (y_t-y_c)	Difficulty difference (y_t-y_c)	1	1846.725*
	Outcome Condition	4	14.147*
	Sample Size	3	32.804*
	Pre-test Length	2	1.744
	Residual		

Note: * $p < 0.05$

Quantifying the Treatment Difference Once again, the next step in the comparison of these three causal inference methods is to calculate the estimated treatment effect for the IRT method and compare this to the estimates from the other methods. Recall from the analysis in the ability-based scenario that the area between the item response curves for the treatment and control group (the differences of these two) represents the treatment effect (example shown in Figure 9). Once again, integration can be used to quantify this treatment effect. Under the cut-score scenario, the decision variable is not included in the 2PL IRT model. Therefore, we are assuming that there is no difference in the distribution in the treatment and control groups.

Mathematically the difference (*Δ*(*u*)) between the two curves can be found by using the following formula. Here a_{yc} represents the estimated discrimination of the outcome for the control group, a_{yt} represents the estimated discrimination of the outcome for the treatment group, b_{yc} represents the estimated difficulty of the outcome for the control group, b_{yt} represents the estimated difficulty of the outcome for the treatment group, and u represents a point on the ability scale.

$$\Delta(u) = \frac{-1}{1+e^{-a_{yc}(u-b_{yc})}} + \frac{-1}{1+e^{-a_{yt}(u-b_{yt})}}$$

Then this difference is integrated, over the conditional distribution of the population of interest, *S*, from negative infinity to positive infinity to quantify the treatment effect, *eff* (*S*).

$$eff(S) = \int_{-\infty}^{\infty} \Delta(u) f(u \mid S) du$$

For each of the three methods, this treatment effect was calculated for each replication in each simulation condition. For the No Difference group, the effect should be zero since both treatment and control group found the outcome equally difficult. Both negative outcome conditions (Low - and High -) should be negative, indicating that there is a negative treatment effect because the control group finds the outcome easier than the treatment group. This value should be larger in magnitude for the High - outcome conditions since a greater difference in difficulty exists initially between groups. The positive outcome conditions (Low + and High +) should both show positive effects since the treatment group found the outcome measure easier than the control group. This value should be larger in magnitude for the High + outcome conditions since there was a greater difference in difficulty initially between groups. Table 19 through Table 22 show the means and standard deviations by method for the treatment effect.

The true values were calculated for each condition. Using the simulating values for the outcome conditions the true treatment effect for each outcome condition is as follows—No Difference: treatment effect = 0; Low -: treatment effect = -1; Low +: treatment effect = 1; High -: treatment effect = -2; and High +: treatment effect = 2. The estimates produced by the IRT method are much larger than the true values when the sample size is small (i.e. N=500), but improve as the sample size gets larger and as the test length increases. Also, the standard deviations are very large indicating that there is great variability of these treatment effect estimates. The LR method is harder to judge from the table.

The estimates from this method do not always follow the expected direction and while the magnitude of the effect is not the same as the true value, it is hard to tell how much they are different. The RD method seems to produce the best treatment estimates with the mean values being as expected in direction and similar in magnitude to the true value.

Table 19

Means (and standard deviations) of treatment effect estimates, by method, for N = 500.

X	OC	IRT	LR	RD
10	No Difference	-13.480 (49.317)	0.107 (1.010)	0.076 (0.508)
	Low -	-21.722 (62.986)	-0.798 (2.054)	-0.871 (0.790)
	Low +	-9.339 (39.850)	0.884 (0.883)	0.956 (0.476)
	High -	-36.395 (85.016)	-1.506 (3.289)	-1.754 (1.148)
	High +	-0.134 (23.810)	1.886 (0.864)	1.920 (0.486)
20	No Difference	-8.328 (38.148)	-0.146 (0.349)	0.133 (0.475)
	Low -	-19.596 (60.655)	0.372 (0.380)	-0.848 (0.579)
	Low +	-6.087 (34.853)	-0.652 (0.329)	1.121 (0.454)
	High -	-33.782 (84.523)	1.035 (0.475)	-1.826 (0.706)
	High +	0.248 (16.768)	-1.112 (0.351)	2.120 (0.466)
50	No Difference	-3.281 (22.520)	-0.935 (2.885)	0.168 (0.427)
	Low -	-7.658 (37.658)	-1.616 (5.103)	-0.801 (0.462)
	Low +	-0.789 (16.621)	-0.851 (0.731)	1.196 (0.418)
	High -	-19.212 (62.416)	-3.550 (8.224)	-1.878 (0.547)
	High +	1.723 (5.087)	-1.115 (0.424)	2.206 (0.401)

Note: True values for the outcome conditions are No Difference = 0.000, Low - = -1.000, Low + = 1.000, High - = -2.000, High + = 2.000.

Table 20

Means (and standard deviations) of treatment effect estimates, by method, for N = 1,000.

X	OC	IRT	LR	RD
10	No Difference	-4.722 (27.111)	0.012 (0.682)	0.042 (0.353)
	Low -	-9.573 (39.802)	-0.940 (0.817)	-0.912 (0.419)
	Low +	-1.077 (16.137)	0.874 (0.611)	0.941 (0.331)
	High -	-15.578 (51.996)	-1.733 (0.998)	-1.792 (0.493)
	High +	1.476 (10.088)	1.828 (0.606)	1.896 (0.342)
20	No Difference	-2.946 (21.071)	-0.168 (0.251)	0.125 (0.344)
	Low -	-4.940 (26.811)	0.407 (0.276)	-0.848 (0.380)
	Low +	0.272 (9.659)	-0.656 (0.228)	1.100 (0.325)
	High -	-13.690 (48.804)	1.017 (0.341)	-1.806 (0.468)
	High +	1.746 (4.409)	-1.104 (0.232)	2.091 (0.326)
50	No Difference	-0.592 (7.746)	-0.449 (0.384)	0.167 (0.295)
	Low -	-2.276 (12.607)	-0.191 (1.169)	-0.848 (0.330)
	Low +	0.929 (0.641)	-0.778 (0.314)	1.198 (0.289)
	High -	-6.039 (27.349)	-0.666 (4.573)	-1.833 (0.373)
	High +	1.994 (0.188)	-1.096 (0.283)	2.225 (0.292)

Note: True values for the outcome conditions are No Difference = 0.000, Low - = -1.000, Low + = 1.000, High - = -2.000, High + = 2.000.

Table 21

Means (and standard deviations) of treatment effect estimates, by method, for N = 5,000.

X	OC	IRT	LR	RD
10	**No Difference**	-0.040 (0.299)	0.004 (0.313)	0.038 (0.161)
	Low -	-1.096 (0.469)	-0.927 (0.335)	-0.899 (0.170)
	Low +	0.975 (0.159)	0.894 (0.271)	0.959 (0.148)
	High -	-2.175 (0.816)	-1.840 (0.422)	-1.828 (0.207)
	High +	2.002 (0.093)	1.832 (0.262)	1.900 (0.145)
20	**No Difference**	-0.056 (0.299)	-0.162 (0.111)	0.114 (0.153)
	Low -	-1.098 (0.543)	0.402 (0.123)	-0.864 (0.165)
	Low +	0.986 (0.169)	-0.661 (0.101)	1.105 (0.140)
	High -	-2.239 (1.521)	1.019 (0.143)	-1.848 (0.199)
	High +	1.992 (0.094)	-1.108 (0.106)	2.083 (0.141)
50	**No Difference**	-0.035 (0.209)	-0.391 (0.153)	0.158 (0.128)
	Low -	-1.062 (0.375)	-0.009 (0.192)	-0.844 (0.146)
	Low +	0.989 (0.114)	-0.739 (0.140)	1.176 (0.128)
	High -	-2.121 (0.618)	0.457 (0.242)	-1.848 (0.165)
	High +	1.993 (0.082)	-1.063 (0.129)	2.190 (0.130)

Note: True values for the outcome conditions are No Difference = 0.000, Low - = -1.000, Low + = 1.000, High - = -2.000, High + = 2.000.

Table 22

Means (and standard deviations) of treatment effect estimates, by method, for N = 10,000.

X	OC	IRT	LR	RD
10	**No Difference**	-0.021 (0.185)	-0.021 (0.207)	0.028 (0.108)
	Low -	-1.057 (0.298)	-0.945 (0.236)	-0.901 (0.122)
	Low +	0.992 (0.105)	0.900 (0.193)	0.958 (0.108)
	High -	-2.083 (0.472)	-1.862 (0.286)	-1.836 (0.144)
	High +	2.002 (0.066)	1.834 (0.187)	1.900 (0.104)
20	**No Difference**	-0.028 (0.199)	-0.163 (0.073)	0.114 (0.107)
	Low -	-1.051 (0.323)	0.396 (0.087)	-0.868 (0.119)
	Low +	0.979 (0.113)	-0.671 (0.076)	1.093 (0.098)
	High -	-2.066 (0.482)	1.020 (0.102)	-1.839 (0.138)
	High +	1.993 (0.066)	-1.114 (0.075)	2.079 (0.103)
50	**No Difference**	-0.014 (0.141)	-0.385 (0.108)	0.163 (0.094)
	Low -	-1.029 (0.222)	0.008 (0.131)	-0.847 (0.098)
	Low +	0.991 (0.078)	-0.732 (0.096)	1.172 (0.088)
	High -	-2.028 (0.355)	0.489 (0.167)	-1.842 (0.116)
	High +	1.996 (0.057)	-1.062 (0.088)	2.190 (0.092)

Note: True values for the outcome conditions are No Difference = 0.000, Low - = -1.000, Low + = 1.000, High - = -2.000, High + = 2.000.

To visualize the difference between the estimated effect and true value, bias and RMSE calculations were performed. Figure 21 through Figure 23 show plots of the bias of these treatment effects and Figure 24 through Figure 26 show plots of the RMSE. For comparison, within bias and RMSE, these figures are plotted on the same scale. However, it is important to note that the range of values varied by causal inference method. For the IRT method, bias ranged from -34.400 to -0.003 and RMSE from 0.000 to 97.100. For LR, bias ranged from -3.120 to 3.400 and RMSE from 0.000 to 8.370. Finally for RD, bias ranged from -.105 to 0.247 and RMSE from 0.000 to 1.174.

From Figure 21 and Figure 24 it can be seen that the bias and RMSE of the estimates from the IRT method are large, specifically for the two conditions of negative treatment effect (Low - and High -). Although, in all conditions the values approach zero when the sample size is 5,000 and larger. The High + condition for the IRT method shows the best parameter recovery of the condition. From Figure 22 and Figure 25 it can be seen that the LR method is better at estimating the treatment effect than the IRT method, for the cut-score scenario. However, like the IRT method, the LR method is having estimation issues not only in the High - condition, but also in the High + condition, which is seen in the plot of the bias. From the plot of the RMSE, the High - condition seems to be having issues. Yet, as in the IRT method, the estimates get better as the sample size increases and the test length increases. From Figure 23 and Figure 26 it can be seen that the RD method has the best treatment effect estimation of the three methods. The bias and RMSE are both close to zero for all conditions and are much closer than the other methods.

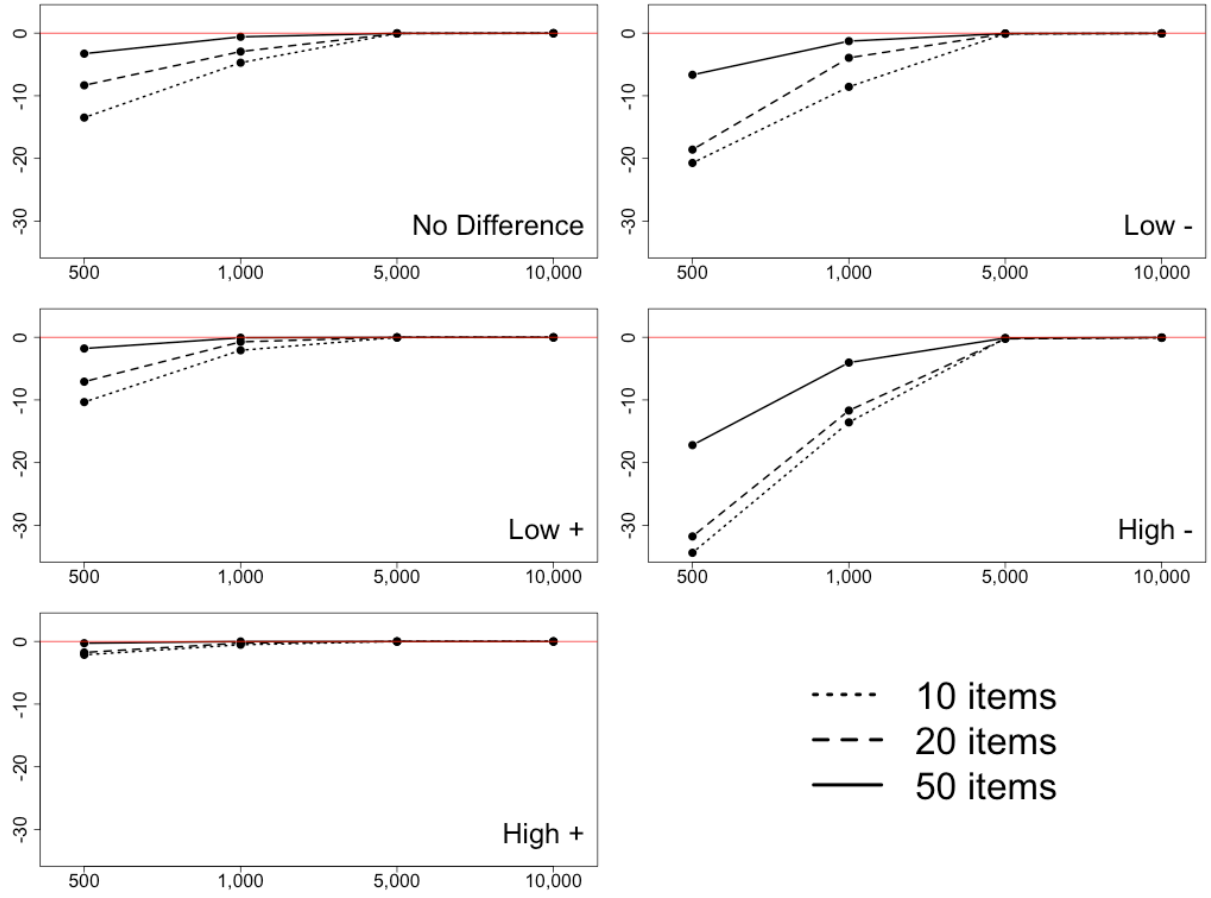

Figure 21: Profile plot of bias by outcome condition for the IRT method under cut-score scenario

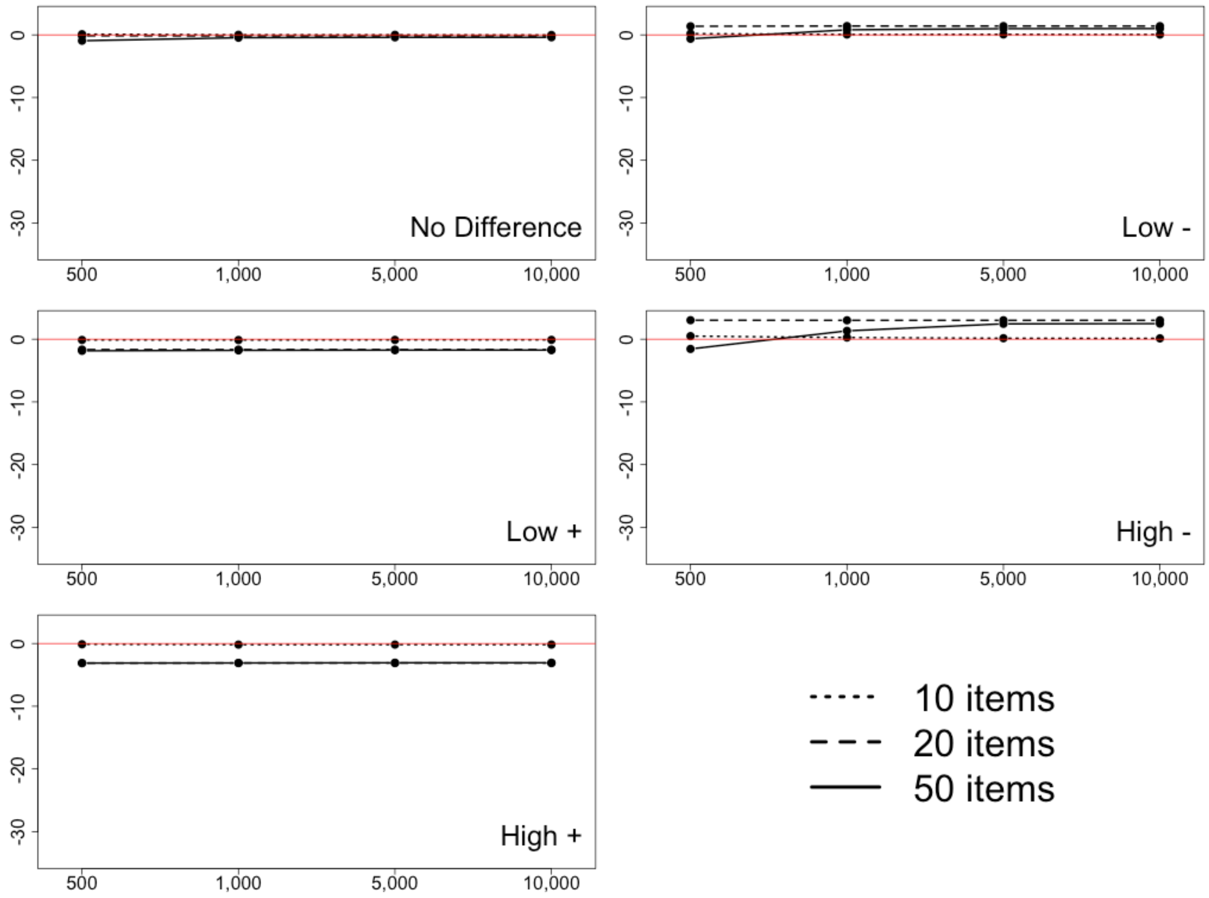

Figure 22: Profile plot of bias by outcome condition for the LR method under cut-score scenario.

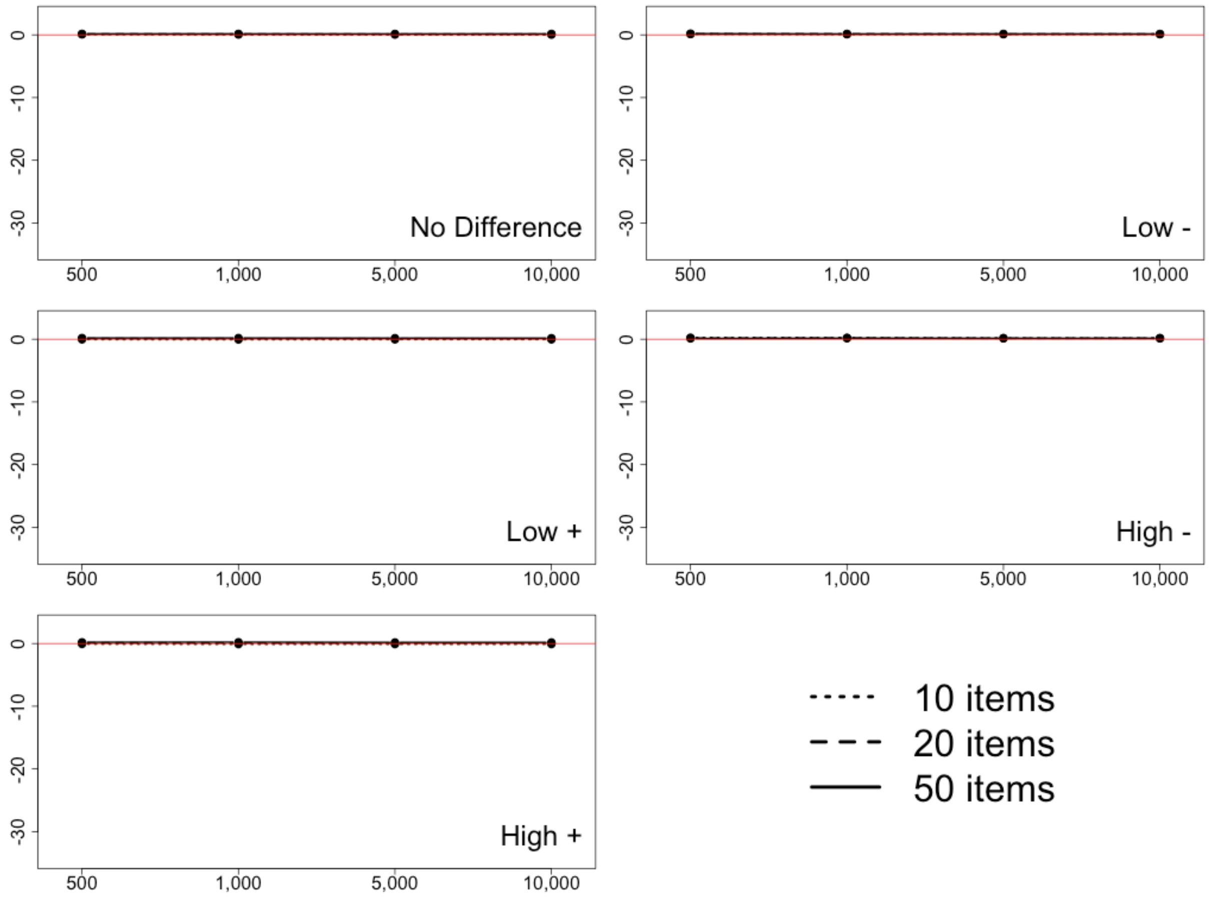

Figure 23: Profile plot of bias by outcome condition for the RD method under cut-score scenario.

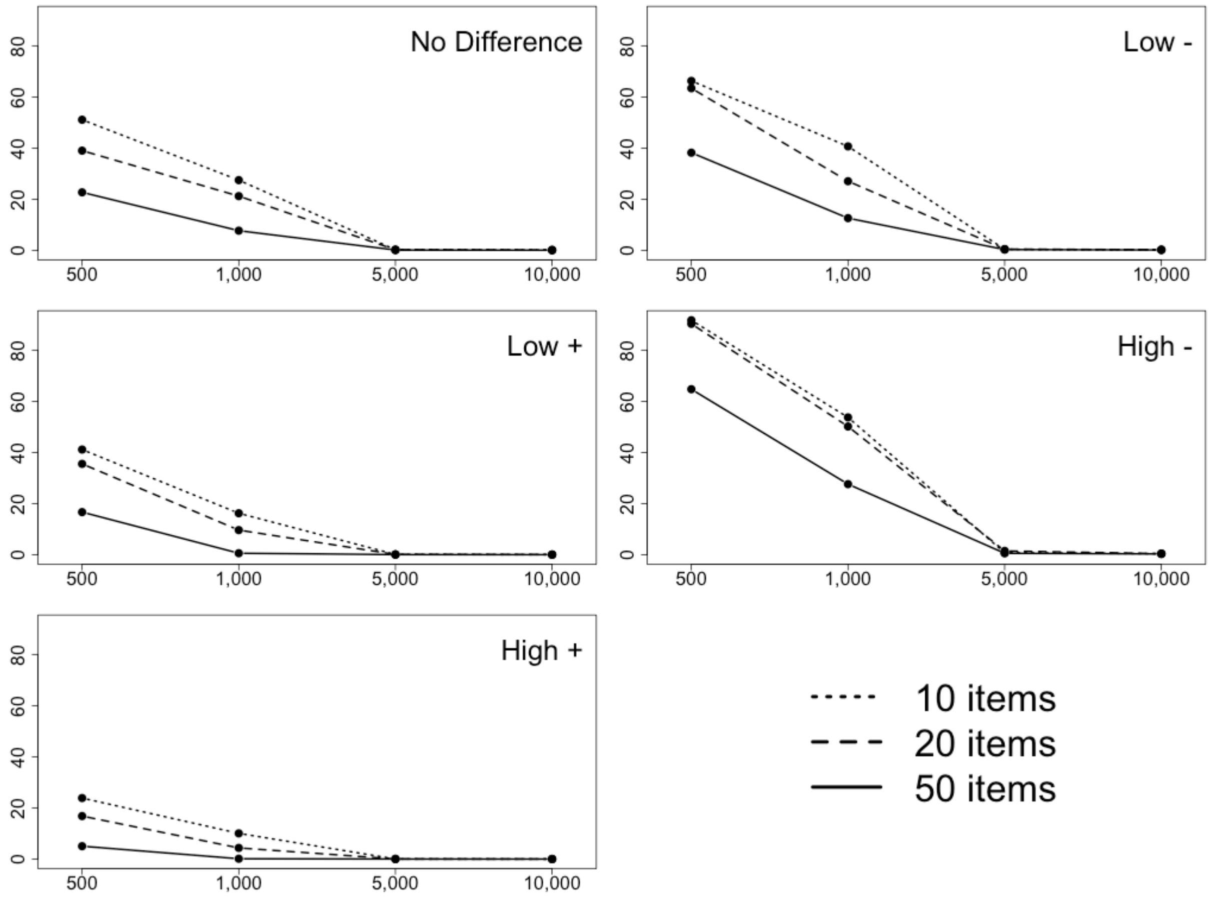

Figure 24: Profile plot of RMSE by outcome condition for the IRT method under cut-score scenario.

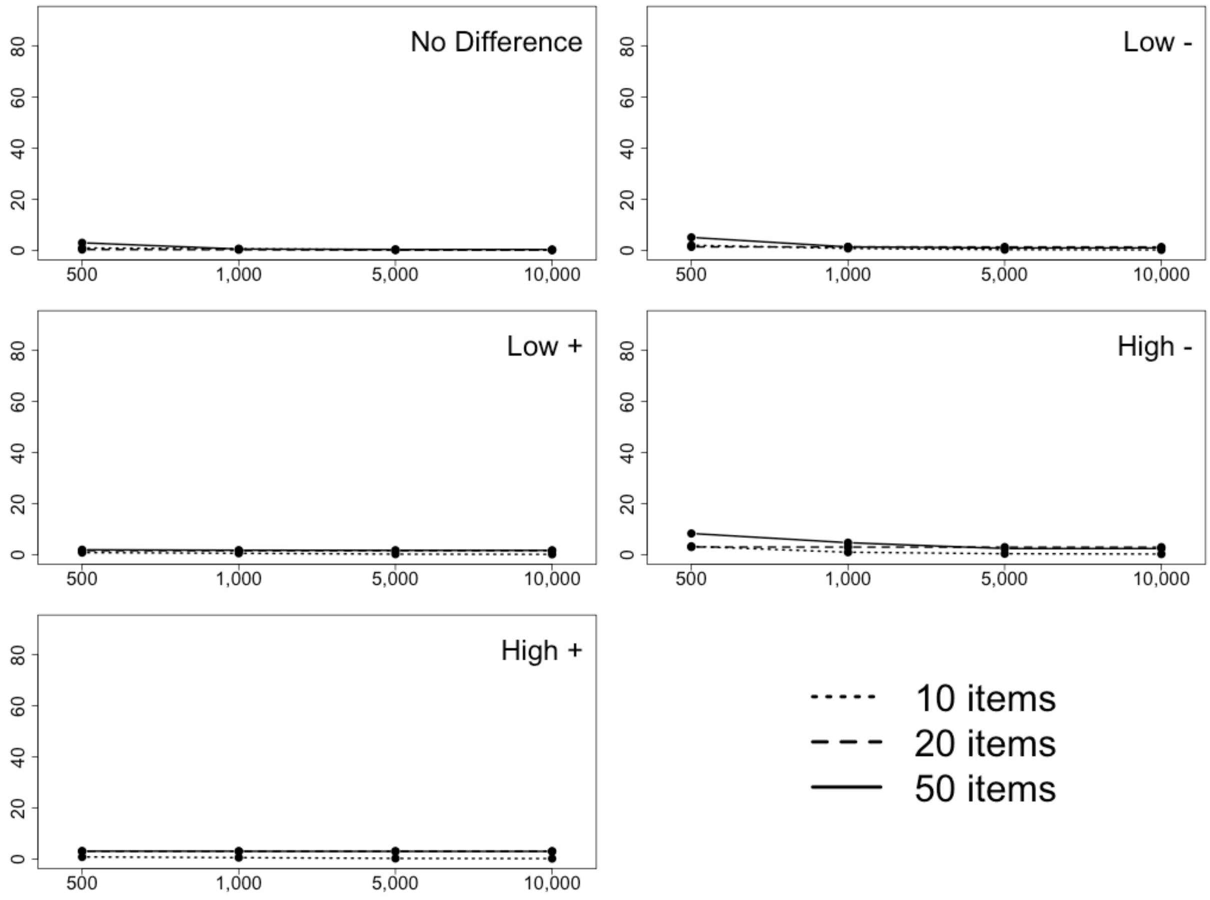

Figure 25: Profile plot of RMSE by outcome condition for the LR method under cut-score scenario.

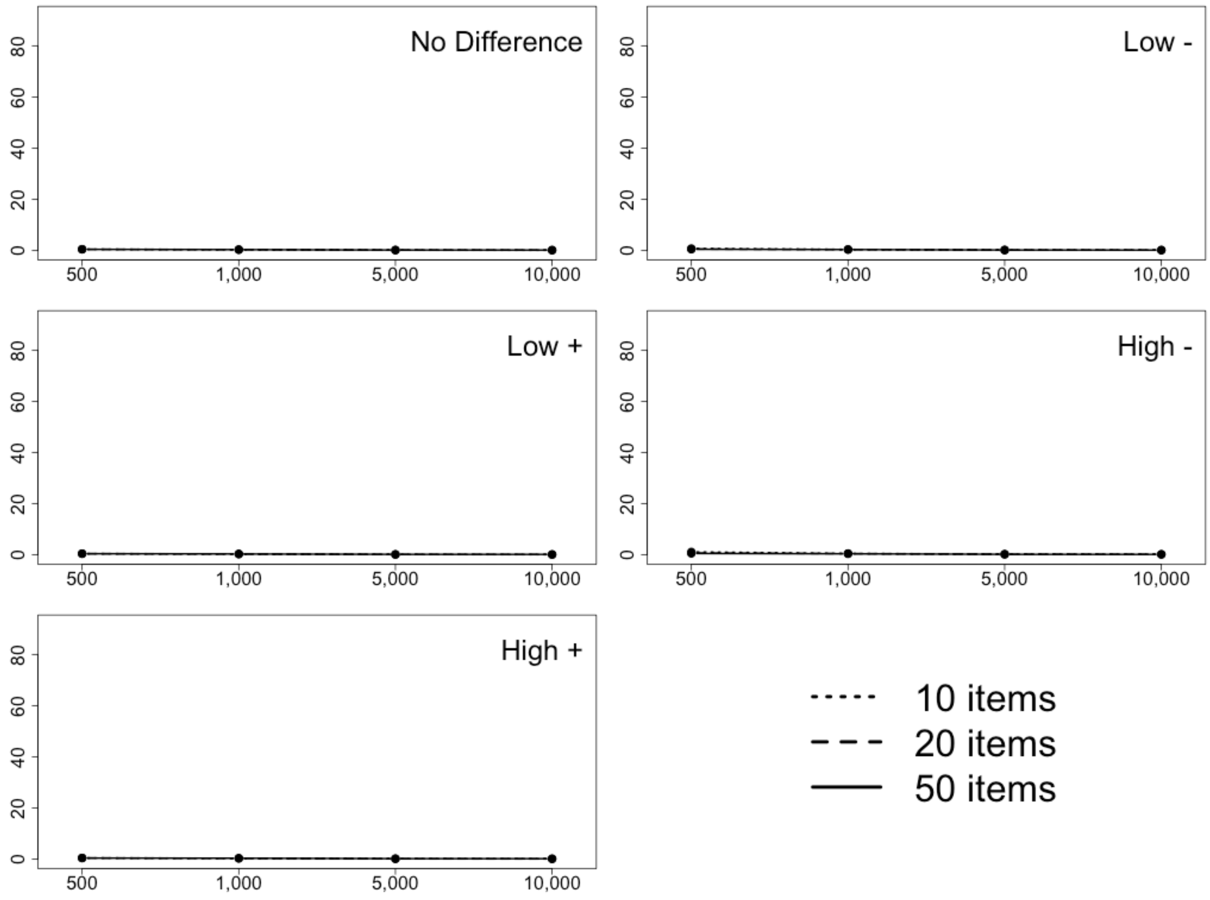

Figure 26: Profile plot of RMSE by outcome condition for the RD method under cut-score scenario.

Real World Data

To examine how this method performs on real world data, item-level data from a large-scale college admission exam, test scores from an advanced end-of-course high school exam, and college major were obtained. To focus in on mathematics ability, only the responses on the three sections of the math portion of the exam were used. Items that were correct were coded as "1." Items that were incorrect were coded as "0." Omitted and not reached items were also coded as "0" because of the estimation methods used in the R function *est()* in the package IRTOYS (Partchev, 2012). The treatment was the decision to take a STEM course, represented by the end-of-course exam, during the senior year and the outcome was deciding to be a STEM major in college.

The sample contained 2,655 students with 1,821 students taking an advanced STEM course during their senior year. The mathematics portion of the test contained 54 items and for this sample, the total score on this portion ranged from 11 to 54 with a mean of 40.012 (SD = 8.238). For the students who took the STEM course during their senior year (the treatment group) the total score ranged from 13 to 54 (M = 42.087, SD = 7.547). For those who did not take the STEM course during the senior year (the control group) the total score ranged from 11 to 53 (M = 35.483, SD = 7.862).

Figure 27 shows the item characteristic curves for both the non-STEM (control, labeled "1") and STEM (treatment, labeled "2") students. The IRT model estimated the following outcome parameters of discrimination (*a*) and difficulty (*b*) for the non-STEM group as *a* = 0.355 and *b* = 4.275 and for the STEM group as *a* = 0.528 and *b* = 1.166. It is shown that over the entire ability scale, the STEM students have a higher probability of becoming a STEM major in college than those who did not take a STEM exam during the senior year. Also, as ability estimates get larger, the difference between the probabilities of the two groups majoring in STEM fields in college increases. This difference can be quantified by integrating to find the area between the two curves. To be meaningful, and comparable to the LR and propensity scores, this integration should be weighted by the ability distribution of those in the treatment group. Doing this results in an estimate of the effect of the treatment on the treated (TT). The value of the TT for the IRT model is 0.191. The significance of this value is found by performing a Wald test. For this test, the null hypothesis is that the discrimination in the control group is equal to the discrimination value in the treatment and the difficulty value in the control group is equal to the difficulty value in the treatment group (i.e. the two groups have the same difficulty and discrimination value). The alternative is that at least one pair is different. Here, this value is highly significant (p = 4.28e^{-13}). This indicates that taking the STEM class during the senior year significantly increases the probability of becoming a STEM major in college.

LR and propensity score models were also run on this data. The LR model was centered around the mean of the treatment group. Doing this resulted in a statistically significant effect of the decision to take an STEM exam in the senior year (coef = 1.005 p < 0.001). The interaction between the total score on the math section and the decision to take STEM was not significant at the 0.05 level (coef = 0.0270, p 0.0612), indicating that the main effect of the decision can be interpreted. There is a significant effect of taking an advanced STEM course in the senior year and majoring in a STEM field in college. A plot of the fitted probabilities of majoring in STEM found from the LR analysis and the total score shown in Figure 28. Also, the average predicted probability of becoming a STEM major for those who took the senior year class is .389. If these students did not take the senior year class their average predicted probability of becoming a STEM major is 0.188. The difference of these two, 0.201, is the average effect of the treatment on the treated. The propensity score model also estimated a statistically significant effect of taking an STEM exam during the senior year, effect = 0.195 (p < 0.001), while using the different score response patterns to match students. Based on these two existing methods of causal inference, it can be concluded that there is a statistically significant relationship between taking an STEM exam during the senior year and majoring in a STEM field in college.

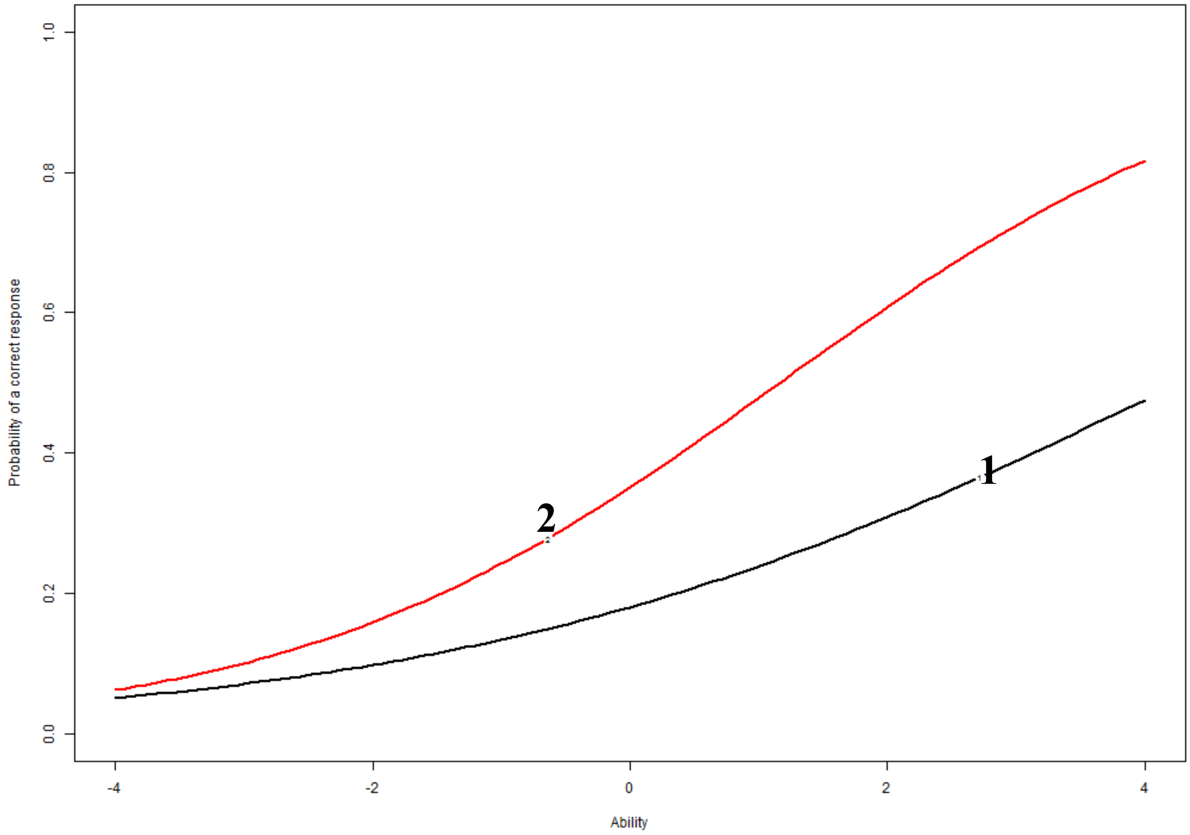

Figure 27: Plot of the item characteristic curve for the outcome variables in the real world data. The curve labeled “1” shows the control, or non-STEM, exam condition and the curve labeled “2” shows the treatment or STEM exam condition.

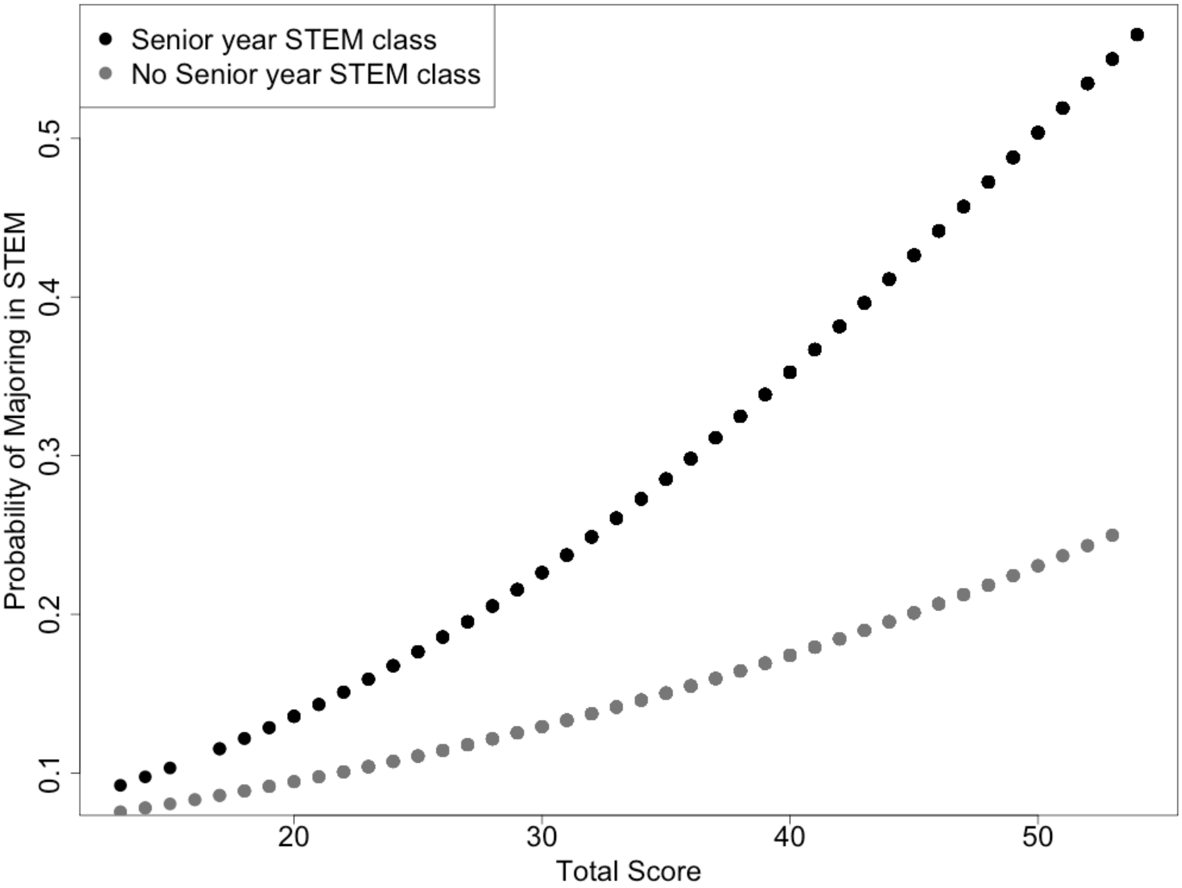

Figure 28: Plot of the fitted values from the LR model by the total test score. The black dots show the fitted probability of majoring in STEM for the group that took the senior year advanced STEM course and the gray dots shows this for the group that did not take the course during their senior year.

Chapter 5: Discussion

Randomized experiments are the gold standard for estimating causal effects. However, in observational research and most educational settings, randomized experiments are hard to come by. To compensate for this lack of randomized "gold-standard" research opportunities, different methods of causal inference have been created. These methods provide some form of correction to make inferences less biased. Some methods match students prior to the intervention, some correct using statistical methods before the analyses, and some set up strict assumptions of what types of data can be used. However, these methods can still provide biased results if the underlying assumptions are not met—which many times they are not.

The goal of this research was to develop and evaluate a new causal inference model based in IRT for use when a pre-intervention test is used. This method was developed to combat many of the currently existing methods shortcomings, specifically the bias associated when a treatment and an outcome are no longer conditionally independent given pre-test performance. To adequately see if this method did combat these shortcomings, it was evaluated under two different scenarios—ability-based and cut-score—and compared to existing methods for causal inference applicable to each scenario. Recall that the cut-score scenario had a treatment decision based on a specified cut-score on the pre-test (see Figure 2) and the ability-based model included a treatment decision that was not directly decided by the pre-test (see Figure 3). The cut-score scenario was identified as a scenario in which existing methods were still applicable, but when the assumption of conditional independence between decision and outcomes, given the pre-test score, no longer held (i.e. the under the ability-based scenario), these existing methods would not be appropriate. Under the ability-based scenario, existing methods would have trouble estimating accurate treatment effects, since a key assumption was missing, and bias would enter the analyses.

Under the ability-based scenario the IRT method outperformed the existing methods in all method comparisons. In terms of Type I error, the IRT method had excellent rates, demonstrating that the method would not identify differences if they did not exist. The existing methods had larger Type I error rates indicating that both LR and propensity score identified more differences when they did not actually exist than the IRT method. In fact, LR and propensity score were struggling significantly at times in terms of Type I error rates. Also, the IRT method showed excellent power of detecting true differences. Even when LR and propensity score methods had good power, the IRT method showed better results. This indicates that the IRT method is better than the existing methods to tell when true differences between the treatment and control group exist and when they do not exist and it is also less likely to give false results.

In terms of parameter recovery, the IRT method was able to recovery item difficulty and discrimination parameters well for both treatment and control groups across the various simulation conditions. The parameter recovery improved as the sample size increased. This provides strong evidence towards the accuracy of parameter estimation under the IRT method. This is very important since the treatment effect is found using the item parameters from the outcome variables. So, not only will the IRT method accurately identify differences when they do exist, but the parameters used to quantify the differences are extremely accurate.

Most notable for the IRT method, under the ability-based scenario, is its ability to actually estimate treatment effects over all 60 simulation conditions. Furthermore, it is not just the accuracy of the treatment effect estimates, but the fact that the IRT method performed far superior to the existing methods while estimating treatment effects. The IRT method produced better, more accurate, and less biased estimates of the treatment effect than the very popular and widely used propensity score matching method.

The IRT method was run through the gamut, under the ability-based scenario, and came out victoriously. Not only did the method demonstrate that it could perform well, yet it also showed that it could combat the biases that other existing causal inference methods would fall victim to when the assumption of conditional independence was broken. Less biased estimates of treatment effects were demonstrated numerous times for all pretest lengths, sample sizes, and outcome conditions which is very important since estimates can be either positively or negatively biased.

Next, the IRT method was investigated under the cut-score scenario. Recall once more that this scenario is one in which the assumptions of currently existing methods still hold. Under this scenario the IRT method was compared to LR and RD methods in terms of Type I and Type II error, parameter recovery ability, and treatment effect estimation. The IRT method had better Type I error rates than the other two methods. This is very important since a researcher does not want evidence of differences to be found unless those differences actually exist. In terms of power, the IRT method, LR and RD had good power throughout all the simulation conditions, however the IRT method struggled for smaller sample sizes.

In terms of parameter recovery, when the sample size was 5,000 and above, the IRT method performed better than LR and comparable to RD, in terms of bias and RMSE. The High - condition is where the IRT method performed poorest for sample sizes of 500 and 1,000 yet, for these sample sizes performed much better in the High +. In terms of estimating the treatment effect, RD produced the most accurate estimates as compared to the other methods.

Once again, IRT produced better estimates for larger sample sizes than it did for smaller ones. LR was not accurate with producing these estimates.

As mentioned previously, the cut-score scenario is one of the scenarios that existing methods of causal inference could handle, as long as they accounted for the previously discussed assumptions and biases. RD is one of the methods that is most appropriate for this type of scenario, which might be evidence for why it performed well. However, it is also of interest to note that the measurement error involved with a pre-test can act almost as a covariate to treatment placement, which under the “fuzzy design” RD can influence the treatment effect estimates and make it more like a randomized experiment and influence the estimates (Hahn, Todd, & VanDerKlaauw, 2001). It could also be likely that the RD method performed well due to measurement error within the pre-test affecting the estimates of the effects. Although the IRT method was not the best method for all aspects of the simulation conditions under the cut-score, it did perform well under larger sample sizes (5,000 and 10,000), longer length tests, and outcome conditions where the difference in difficulty between the treatment and control groups were not exceedingly large (No Difference, Low -, Low +).

The real data that was used to test each of these methods resembles scenarios used in the simulation study. The scenario that this data falls under is the ability-based scenario since scoring a certain total score on the exam does not place a student into the advanced STEM course. Although the sample size of the real data was not studied in these analyses, it does fall above the threshold of 1,000 where issues were seen. Also, the pre-test is near the largest size used in simulation. Together this indicates that estimates found by any of the methods should perform well. As was seen in the analyses, this was the case. Each of the tested methods was able to identify a significant treatment effect and all of the estimates were in the same range as each other. Coupling that fact with the evidence that the IRT provided more accurate estimates of the treatment effect during simulation, one may infer that the treatment effect estimated by the IRT method is the most accurate one.

These analyses provide evidence that the IRT method is very useful for scenarios where a pre-test, a treatment decision, and outcomes are all related to ability, like that in Figure 3. An ability-based scenario is quite common when a pre-test exists because not all decisions are based strictly on test scores. Retention and promotion of students do not rely solely on a test score, nor does placement in an honors or accelerated program. Tests are used as tools to help gauge understanding, comprehension, learning, and ability, but are not the be-all and end-all in decision-making (nor should they ever be). They are a tool to facilitate progress and decisions and must be evaluated as such. The IRT method does just that—it does not put un-do emphasis on one test score, but takes into account other factors associated with overall ability. Because of this, this method produces less biased and more accurate estimates of treatment effects. In its most general terms, this IRT method can be thought of as an ANCOVA that matches on true ability, yet it is all model based. The IRT method is powerful in its ability to gather information about each group member by using that member's pre-test, decision, and outcome behavior to be able and estimate their true ability. Furthermore, within the method this estimate of true ability is used to match between treatment and control groups, allowing model-based matches that control for extraneous measurement error and unobserved confounding covariates.

While the IRT method can be used in an array of settings, there are certain data requirements for its use. First, the test data must be item-level test response data. This means that data for the correctness of each question must be available. The data should come from a high-quality, reliable test, if possible, since a poor test will not provide accurate information about the student. Next, the decision variable needs to be an ordinal categorical variable with two or more levels. If the decision has more than two levels, a polytomous IRT model, like the GPCM, must be used. Finally, the outcome variable that is used in the IRT model must be input as a ordinal categorical variable. If the outcome is continuous, it should be broken down into categories prior to being input into the model. For example, if the outcome is GPA, it should be broken down into categories of interest (e.g. B or higher and lower than a B) before including it into the model.

Limitations

As briefly mentioned earlier, the IRT model struggled when applied to smaller sample sizes under the cut-score scenario. This struggle can be attributed to how the treatment and control groups were populated under the cut-score scenario. A cut-score equal to half the items on the test was used, but this caused the groups to become uneven due to the difficulty of the pre-test items and the ability of the simulated group members. Certain replications within the simulation conditions with smaller sample sizes produced very large difficulty estimates for either treatment group—indicating that all members got the question correct or all members got the question wrong.

Figure 29 illustrates the outcome behavior one of these data sets. This figure shows the item characteristic curves (ICC) for the treatment and control groups. These curves show the probability of success on the outcome across the entire ability spectrum. This specific data is one of the replications from a simulation of N = 500 on a pre-test of 20 items for the No Difference outcome condition. This figure illustrates that there was no variation in the probability of success for those in the control group. However, both groups have very similar distributions of ability indicating that even though they behave very differently on the outcome, they have similar abilities. This allowed for very little variation within the model, resulting in estimates that behaved poorly. As the sample size increased, estimates behaved better and appear more comparable to the LR and RD methods.

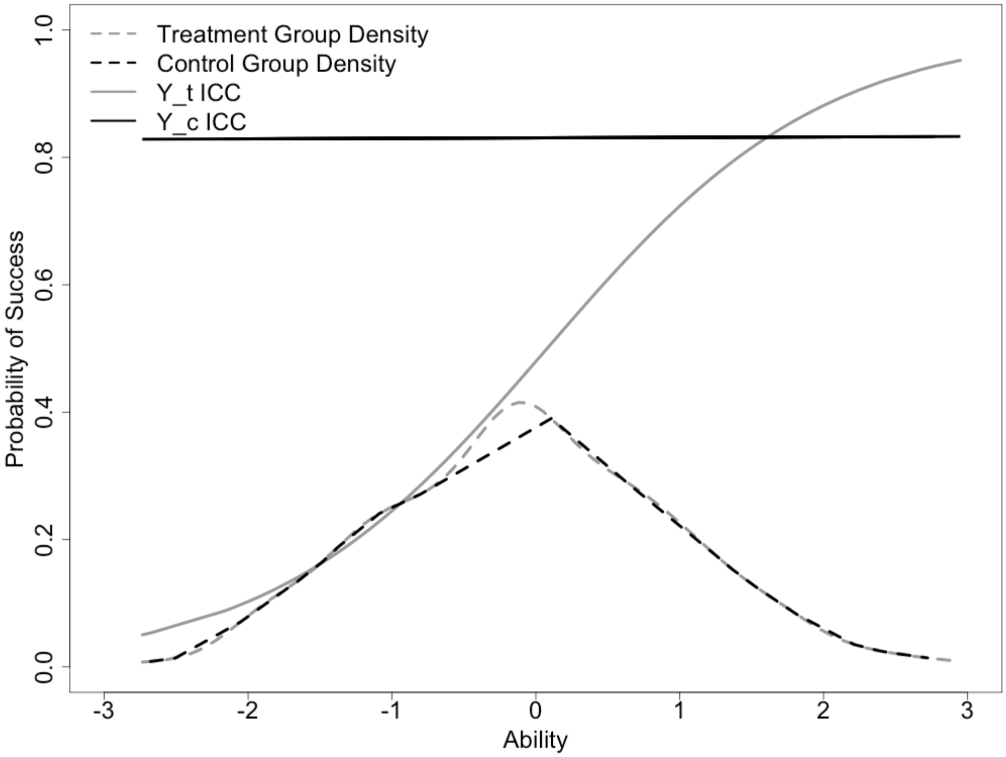

Figure 29: Visualization of treatment and control groups' probability of success and the distribution of each group's abilities.

Another limitation of this study was the use of the *est* ( ) function in the IRTOYS (Partchev, 2012) package in R. As mentioned before, this function called on the ICL program (Hanson, 2002) to fit and estimate the 2PL model on the data. However, the ICL program does not return anything other than item parameter estimates. It does not return standard errors of any type. Because of this, standard errors in the simulations had to be calculated using Monte Carlo approximations of bias and RMSE. Other estimation methods do exist in R, like the LTM package (Rizopoulos, 2011), however the *ltm* ( ) function within this package took between 20 to

40 minutes to run one replication within a simulation condition. Recall that 1,000 replications were performed within each of the 60 simulations conditions. Due to time constraints *ltm* ( ) was not a viable option.

Another limitation of this research is the selection of the difficulty values used in the outcome condition section of the simulation conditions. These were selected at the beginning of the research using arbitrary values. Upon completion of the analysis, it was realized that these values are actually quite large. In fact a difference in difficulty of -0.5 to 0.5, as seen in both of the Low outcome conditions (Low – and Low +), is actually a large effect. It is believed that this limited the amount of variation observed between conditions in analyses such as the Type I and Type II error analysis.

A final limitation of this study was the availability of "real world" data on which to test the IRT method. It was not easy to obtain the data used in this study. In fact, finding item-level pre-test data, with an associated decision, and outcomes was close to impossible. It would be helpful to explore this method using other datasets to truly grasp the usefulness of the method.

Areas for Future Research

One of the most pertinent areas for future research would be to find other applicable data sets and use this model to estimate the effect of a decision. Finding access to the correct data source, like a placement test at a community college, would be very helpful.

Also, further exploration must be done to determine if there is an optimal situation under a cut-score scenario where the IRT method is best used. There seems to be multiple factors that affect how the IRT method performs under the cut-score model. The two most obvious are the determination of the cut-score and the sample size. These two factors are intertwined, as well. It is possible that in smaller samples, the distribution of subjects into treatment and control created very little variation within groups, making estimation difficult. So, an optimal cut-score should be explored by trying different percentages of subjects in both treatment and control groups. Then the model should be tested under the cut-score method against existing methods to see how it performs.

This study set the outcome discrimination to one. Future studies should explore what happens when the outcome discrimination is no longer held constant at one. This segues into more future research as well. Different types of IRT models, like the 3PL model or GPCM, should be explored as well. The transition to GPCM would be an easy one since the GPCM allows for polytomous data; no longer would the test, decision, and outcome have to be dichotomous. This would allow for decisions with multiple levels (e.g. studying the effect of being place in remedial, college prep, or honors English) and multiple level outcomes (e.g. failure, moderate success, success). Incorporating a different IRT model would allow for new situations to be modeled by this method.

Also, it is important to see how the method behaves when covariates are added and how that compares to existing methods. Covariates could be added to any part of the model. They could be associated with the pretest, the decision, the outcome, or with ability itself. These components could even be interrelated, covarying with each other. The addition of covariates would require the use of multiple group IRT models (Muthén & Christoffersson, 1981; Muthén & Lehman, 1985). More complex situations need to be explored to fully understand the scope of the application of the IRT method.

Finally, it would be interesting to see how the IRT method could be combined with Cognitive Diagnostic Models (CDM; Tatsuoka & Tatusoka, 1992). CDMs are latent trait models that assess the presence or absence of skills. They have been used within a higher dimensional attribute space to specify the joint distribution of the latent variables (de la Torre & Douglas, 2004). It would be interesting to see how CDMs could be combined into the IRT method. Perhaps the IRT method can be tied back to the skill sets that students with identical abilities possess. If so, it would be highly interesting to investigate the effectiveness of the treatment placement in the realm of the possessed skills. A more concrete example of where this might be applicable is the following. Imagine that a group of students took a college admissions placement test, given by the school itself. Then the student is placed into a level of a first-year class (e.g. low, medium, high). The outcome could be the level of the student's final grade or the progressing to the next course level. It would be useful to evaluate the level of placement in terms of the skills that student possesses, based on the placement test. This would seem to have implications for college retention and completion rates, among other uses.

This dissertation has illustrated the usefulness of approaching a causal inference model from an IRT perspective. It has shown the statistical properties of this IRT method in two different scenarios and compared it to current methods. In addition, resources have been provided to apply this new method to future studies in various fields.

Reference List

Abadzi, H. (1984). Ability grouping effects on academic achievement and self-esteem in a southwestern school district. *Journal of Educational Research, 77*(5), 287-292.

Abadzi, H. (1985). Ability grouping effects on academic achievement and self-esteem: Who performs in the long run as expected. *Journal of Educational Research, 79*(1), 36-40.

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association, 91*(434), 444-455. doi: 10.1080/01621459.1996.10476902

Angrist, J. D. & Lavy, V. (1999) Using maimonides' rule to estimate the effect of class size on scholastic achievement. *Quarterly Journal of Economics, 114*(2), 533-576. doi: 10.1162/003355399556061

Belin, T. R. & Normand, S. T. (2009). The role of ANCOVA in analyzing experimental data. *Psychiatric Annals, 39*(7), 753-759. doi: 10.3928/00485713-20090625-01

Bifulco, R. (2010). *Can propensity score analysis replicate estimates based on random assignment in evaluations of school choice? A within-study comparison* (Center for Policy Research Working Paper 124). Syracuse, NY: Center for Policy Research. Retrieved March 30, 2011 from http://cpr.maxwell.syr.edu/cprwps/pdf/wp124.pdf doi: 10.2139/ssrn.1805890

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord and M.R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397-479). Reading, MA: Addison-Wesley.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*(1), 29-51. doi: 10.1007/BF02291411

Brand, J. E. & Xie, Y. (2010). Who benefits most from college? Evidence for negative selection in heterogeneous economic returns to higher education. *American Sociological Review, 75*(2), 273-302. doi: 10.1177/0003122410363567

Briggs, D. C. (2004). Causal inference and the heckman model. *Journal of Educational and Behavioral Statistics, 29*(4), 397-420. doi: 10.3102/10769986029004397

Casella, G. and Berger, R. L. (2002). *Statistical Inference*. California: Duxbury.

College Board (2013). *Student—Frequently Asked Questions: ACCUPLACER*. http://accuplacer.collegeboard.org/students/faqs

Crone, D. A., Stoolmiller, M., Baker, S. K., & Fien, H. (2012, Fall). *The Middle School Intervention Project: Use of a Regression Discontinuity Design to Evaluate a Multi-Component Intervention for Struggling Readers in Middle School in Six School Districts.* Society for Research on Educational Effectiveness, Conference, Washington, DC.

de la Torre, J. & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika, 69*(3), 333-353. doi: 10.1007/BF02295640

Fan, X. & Nowell, D. L. (2011). Using propensity score matching in educational research. *Gifted Child Quarterly, 55*(1), 74-79. doi: 10.1177/0016986210390635

Gonzales, P., Williams, T., Jocelyn, L., Roey, S., Kastberg, D., and Brenwald, S. (2008). *Highlights from TIMSS 2007: Mathematics and Science Achievement of U.S. Fourth- and Eighth-Grade Students in an International Context* (NCES 2009-001). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

Hahn, J., Todd, P., & Van der Klaauw, W. (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica, 69*(1), 201-209. doi: 10.1111/1468-0262.00183

Hanson, B. A. (2002). *ICL: IRT Command Language.* www.b-a-h.com

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica, 47*(1), 153-161. doi: 10.2307/1912352

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81*(396), 945-960. doi**:** 10.1080/01621459.1986.10478354

Holland, P. W. & Rosenbaum, P. R. (1986). Condition association and unidimensionality in monotone latent variable models. *Annals of Statistics 14*(4), 1523-1543. doi: 10.1214/aos/1176350174

Hong, G. & Raudenbush, S. W. (2005). Effects of kindergarten retention policy on children's cognitive growth in reading and mathematics. *Educational Evaluation and Policy Analysis, 27*(3), 205-244. doi: 10.3102/01623737027003205

Hong, G. & Raudenbush, S. W. (2006). Evaluating kindergarten retention policy: A case study of causal inference for multilevel observational data. *Journal of the American Statistical Association, 101*(475), 901-910. doi: 10.1198/016214506000000447

Horn, C., McCoy, Z., Campbell, L., & Brock, C. (2009). Remedial testing and placement in community colleges. *Community College Journal of Research and Practice, 33*, 510-526. doi: 10.1080/10668920802662412

Imbens, G. W. & Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics, 142*, 615-635. doi: 10.1016/j.jeconom.2007.05.001

Lee, D. S. & Card, D. (2008). Regression discontinuity inference with specification error. *Journal of Econometrics, 142*, 655-674. doi: 10.1016/j.jeconom.2007.05.003

Lee, D. S. & Lemieux, T. (2009). *Regression discontinuity designs in econometrics* (NBER Working Paper 14723). Cambridge, MA: National Bureau of Economic Research. Retrieved October 1, 2010 from http://www.nber.org/papers/w14723

Legislative Analyst's Office (2011, March) *Are Entering Freshmen Prepared for College-level Work?* (Issue brief no. 2). Retrieved from the Legislative Analyst's Office Higher Education section website: http://www.lao.ca.gov/sections/higher_ed/FAQs/Higher_Education_Issue_02.pdf

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

Matthews, M. S., Peters, S. J., & Housand, A. M. (2012) Regression discontinuity design in gifted and talented education. *Gifted Child Quarterly, 56*(2), 105-112. doi: 10.1177/0016986212444845

Melguizo, T. (2010). Are students of color more likely to graduate from college if they attend more selective institutions? Evidence from a cohort of recipients and nonrecipients of the gates millennium scholarship program. *Educational Evaluation and Policy Analysis, 32*(2), 230-248. doi: 10.3102/0162373710367681

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika, 56*(2), 177-196. doi: 10.1007/BF02294457

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement 16*(2), 159-176. doi: 10.1177/014662169201600206

Muraki, E. & Bock, R. D. (1997). *PARSCALE: IRT item analysis and test scoring for rating-scale data.* [Computer program]. Scientific Software International, Chicago, IL.

Muthén, B. & Christoffersson, A. (1981). Simultaneous factor analysis of dichotomous variables in several groups. *Psychometrika, 46*(4), 407-419. doi: 10.1007/BF02293798

Muthén, B. & Lehman, J. (1985). Multiple group IRT modeling: Applications to item bias analysis. *Journal of Educational and Behavioral Statistics 10*(2), 133-142. doi: 10.3102/10769986010002133

Nomi, T. & Allensworth E. (2009). "Double-dose" algebra as an alternative strategy to remediation: Effects on students' academic outcomes. *Journal of Research on Educational Effectiveness, 2*, 111-148. doi: 10.1080/19345740802676739

Ou, D. (2010). To leave or not to leave? A regression discontinuity analysis of the impact of failing the high school exit exam. *Economics of Education Review, 29*, 171-186. doi: 10.1016/j.econedurev.2009.06.002

Owen, A. L. (2010). Grades, gender, and encouragement: A regression discontinuity analysis. *The Journal of Economic Education, 41*(3), 217-234. doi: 10.1080/00220485.2010.486718

Partchev, I. (2012). *Simple Interface to the Estimation and Plotting of IRT Models*. Software package for R. [Computer program]. http://cran.r-project.org/web/packages/irtoys/irtoys.pdf

Pearl, J. (2000). The logic of counterfactuals in causal inference (Discussion of 'Casual inference without counterfactuals' by A.P. Dawid). *Journal of the American Statistical Association, 95*(450), 428-435. (Technical Report, R-269). http://ftp.cs.ucla.edu/pub/stat_ser/R269.pdf

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Nielsen & Lydiche, Copenhagen.

Rizopoulos, D. (2011). *Latent Trait Models under IRT*. Software package for R. [Computer program]. http://rwiki.sciviews.org/doku.php?id=packages:cran:ltm

Rosenbaum, P. R. (1984). Testing the conditional independence and monotonicity assumptions of item response theory. *Psychometrika 49*, 425-436. doi: 10.1007/BF02306030

Rosenbaum, P. R. (2010). *Design of observational studies*. New York: Springer. doi: 10.1007/978-1-4419-1213-8

Rosenbaum, P. R & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*(1), 41-55. doi: 10.1093/biomet/70.1.41

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 66*(5), 688-701. doi: 10.1037/h0037350

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*(3), 581-92. doi: 10.1093/biomet/63.3.581

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics, 6*(1), 34-58. doi: 10.1214/aos/1176344064

Rubin, D. R. (1986). Which ifs have causal answers (Comment on Statistics and Causal Inference by P. W. Holland). *Journal of the American Statistical Association, 81*(396), 961-962. doi: 10.1080/01621459.1986.10478355

Seaver, W. B. & Quarton, R. J. (1976). Regression discontinuity analysis of dean's list effects. *Journal of Educational Psychology, 68*(4), 459-465. doi: 10.1037/0022-0663.68.4.459

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2001). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference.* Boston: Houghton Mifflin.

Stone, R. (1993). The assumptions on which causal inferences rest. *Journal of the Royal Statistical Society. Series B (Methodological)*, *55*(2), 455-466.

Tatsuoka, K. K. & Tatsuoka, M. M. (1992). A psychometrically sound cognitive diagnostic model: Effect of remediation as empirical validity. Research Report. Princeton, NJ: Educational Testing Service.

Thistlethwaite, D. L. & Campbell, D. T. (1960). Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational Psychology, 51*(6), 309-317. doi: 10.1037/h0044319

Todd, P. E. (2006). *Evaluating social programs with endogenous program placement and selection of the treated.* Paper presented at a Conference at the Rockefeller Center in Bellagio, Italy. March 2006.

Zimowski, M., Muraki, E., Mislevy, R., & Bock, D. (n.d.) *BILOG-MG.* [Computer program].Scientific Software International, Chicago, IL.

Appendix

Table A1

True value for the parameters of the items in the pre-tests.

Item	Difficulty	Discrimination	Item	Difficulty	Discrimination
1	0.101	0.651	26	0.626	1.431
2	0.141	0.872	27	-0.060	1.361
3	1.482	1.236	28	0.004	0.965
4	0.514	1.377	29	0.461	0.507
5	1.085	0.524	30	0.640	1.093
6	1.162	1.027	31	0.454	1.496
7	0.832	0.999	32	1.067	1.542
8	-0.131	1.098	33	0.226	0.961
9	0.674	0.747	34	1.473	1.587
10	1.426	0.859	35	0.489	1.505
11	0.631	0.698	36	0.205	1.533
12	0.861	1.193	37	-0.136	1.468
13	1.731	1.256	38	0.613	1.068
14	0.378	1.538	39	0.771	1.288
15	0.306	1.305	40	-0.015	0.481
16	0.804	1.203	41	0.451	0.494
17	0.906	1.121	42	1.114	1.519
18	0.360	1.303	43	-0.463	1.406
19	0.938	1.464	44	-0.080	0.816
20	1.009	1.525	45	-0.721	1.058
21	0.987	0.966	46	0.266	1.185
22	0.746	0.732	47	0.969	0.775
23	-0.053	1.167	48	-0.143	1.105
24	1.610	1.007	49	0.044	0.805
25	1.319	1.257	50	0.647	1.904

Note: Items 1-10 represent the 10-item test; items 1-20 represent the 20-item test; items 1-50 represent the 50-item test.

Table A2

True parameter values and mean IRT estimated parameter values for the outcome in the control and treatment group for the 15 simulations with sample size 500, for the ability-based scenario.

			Difficulty				**Discrimination**			
			True	**Est.**	**Bias**	**RMSE**	**True**	**Est.**	**Bias**	**RMSE**
10 items	**No Difference**	**Treatment**	0	-0.028	-0.028	0.207	1	1.014	0.014	0.258
		Control	0	0.020	0.020	0.201	1	1.023	0.023	0.289
	Low -	**Treatment**	0.5	0.502	0.002	0.152	1	1.032	0.032	0.276
		Control	-0.5	-0.505	-0.005	0.163	1	1.032	0.032	0.289
	Low +	**Treatment**	-0.5	-0.560	-0.060	0.352	1	1.032	0.032	0.289
		Control	0.5	0.532	0.032	0.292	1	1.049	0.049	0.306
	High -	**Treatment**	1	1.027	0.027	0.224	1	1.034	0.034	0.274
		Control	-1	-1.011	-0.011	0.213	1	1.063	0.063	0.301
	High +	**Treatment**	-1	-1.087	-0.087	0.506	1	1.038	0.038	0.316
		Control	1	1.111	0.111	0.659	1	1.029	0.029	0.315
20 items	**No Difference**	**Treatment**	0	-0.008	-0.008	0.168	1	1.026	0.026	0.217
		Control	0	0.010	0.010	0.170	1	1.025	0.025	0.235
	Low -	**Treatment**	0.5	0.495	-0.005	0.152	1	1.016	0.016	0.223
		Control	-0.5	-0.503	-0.003	0.157	1	1.023	0.023	0.231
	Low +	**Treatment**	-0.5	-0.540	-0.040	0.259	1	1.015	0.015	0.228
		Control	0.5	0.525	0.025	0.248	1	1.028	0.028	0.245
	High -	**Treatment**	1	1.016	0.016	0.198	1	1.026	0.026	0.221
		Control	-1	-1.028	-0.028	0.202	1	1.008	0.008	0.231
	High +	**Treatment**	-1	-1.077	-0.077	0.385	1	1.022	0.022	0.255
		Control	1	1.051	0.051	0.374	1	1.033	0.033	0.257
50 items	**No Difference**	**Treatment**	0	-0.004	-0.004	0.156	1	1.031	0.031	0.207
		Control	0	0.002	0.002	0.164	1	1.020	0.020	0.209
	Low -	**Treatment**	0.5	0.503	0.003	0.155	1	1.025	0.025	0.195
		Control	-0.5	-0.506	-0.006	0.153	1	1.018	0.018	0.204
	Low +	**Treatment**	-0.5	-0.527	-0.027	0.237	1	1.019	0.019	0.208
		Control	0.5	0.515	0.015	0.231	1	1.025	0.025	0.213
	High -	**Treatment**	1	1.011	0.011	0.176	1	1.015	0.015	0.196
		Control	-1	-1.017	-0.017	0.185	1	1.024	0.024	0.210
	High +	**Treatment**	-1	-1.037	-0.037	0.328	1	1.029	0.029	0.222
		Control	1	1.032	0.032	0.323	1	1.033	0.033	0.238

Table A3

True parameter values and mean IRT estimated parameter values for the outcome in the control and treatment group for the 15 simulations with sample size 1,000, for the ability-based scenario.

			Difficulty				**Discrimination**			
			True	**Est.**	**Bias**	**RMSE**	**True**	**Est.**	**Bias**	**RMSE**
10 items	**No Difference**	**Treatment**	0	-0.009	-0.009	0.127	1	1.018	0.018	0.192
		Control	0	0.016	0.016	0.129	1	1.014	0.014	0.199
	Low -	**Treatment**	0.5	0.499	-0.001	0.109	1	1.016	0.016	0.181
		Control	-0.5	-0.510	-0.010	0.109	1	1.021	0.021	0.202
	Low +	**Treatment**	-0.5	-0.526	-0.026	0.191	1	1.011	0.011	0.193
		Control	0.5	0.514	0.014	0.183	1	1.028	0.028	0.206
	High -	**Treatment**	1	1.007	0.007	0.135	1	1.012	0.012	0.180
		Control	-1	-1.012	-0.012	0.150	1	1.024	0.024	0.206
	High +	**Treatment**	-1	-1.035	-0.035	0.290	1	1.018	0.018	0.215
		Control	1	1.050	0.050	0.284	1	1.012	0.012	0.207
20 items	**No Difference**	**Treatment**	0	-0.009	-0.009	0.117	1	1.017	0.017	0.156
		Control	0	0.002	0.002	0.114	1	1.014	0.014	0.152
	Low -	**Treatment**	0.5	0.500	0.000	0.103	1	1.018	0.018	0.156
		Control	-0.5	-0.505	-0.005	0.109	1	1.014	0.014	0.167
	Low +	**Treatment**	-0.5	-0.508	-0.008	0.155	1	1.020	0.020	0.165
		Control	0.5	0.520	0.020	0.163	1	1.002	0.002	0.164
	High -	**Treatment**	1	0.999	-0.001	0.127	1	1.021	0.021	0.160
		Control	-1	-1.009	-0.009	0.142	1	1.011	0.011	0.165
	High +	**Treatment**	-1	-1.038	-0.038	0.245	1	1.008	0.008	0.178
		Control	1	1.013	0.013	0.227	1	1.025	0.025	0.177
50 items	**No Difference**	**Treatment**	0	0.000	0.000	0.118	1	1.012	0.012	0.138
		Control	0	-0.001	-0.001	0.117	1	1.011	0.011	0.138
	Low -	**Treatment**	0.5	0.493	-0.007	0.107	1	1.002	0.002	0.135
		Control	-0.5	-0.501	-0.001	0.104	1	1.012	0.012	0.137
	Low +	**Treatment**	-0.5	-0.517	-0.017	0.160	1	1.006	0.006	0.152
		Control	0.5	0.504	0.004	0.148	1	1.011	0.011	0.149
	High -	**Treatment**	1	1.003	0.003	0.127	1	1.009	0.009	0.137
		Control	-1	-1.003	-0.003	0.130	1	1.016	0.016	0.146
	High +	**Treatment**	-1	-1.027	-0.027	0.214	1	1.016	0.016	0.159
		Control	1	1.020	0.020	0.208	1	1.003	0.003	0.151

Table A4

True parameter values and mean IRT estimated parameter values for the outcome in the control and treatment group for the 15 simulations with sample size 5,000, for the ability-based scenario.

			Difficulty				**Discrimination**			
			True	**Est.**	**Bias**	**RMSE**	**True**	**Est.**	**Bias**	**RMSE**
10 items	**No Difference**	**Treatment**	0	-0.005	-0.005	0.054	1	1.001	0.001	0.079
		Control	0	-0.002	-0.002	0.055	1	1.008	0.008	0.084
	Low -	**Treatment**	0.5	0.499	-0.001	0.048	1	1.005	0.005	0.080
		Control	-0.5	-0.504	-0.004	0.047	1	1.006	0.006	0.086
	Low +	**Treatment**	-0.5	-0.503	-0.003	0.076	1	1.009	0.009	0.085
		Control	0.5	0.502	0.002	0.077	1	1.004	0.004	0.086
	High -	**Treatment**	1	1.001	0.001	0.061	1	1.001	0.001	0.080
		Control	-1	-0.999	0.001	0.061	1	1.006	0.006	0.090
	High +	**Treatment**	-1	-1.012	-0.012	0.114	1	1.000	0.000	0.092
		Control	1	1.009	0.009	0.113	1	1.002	0.002	0.093
20 items	**No Difference**	**Treatment**	0	-0.001	-0.001	0.049	1	1.004	0.004	0.067
		Control	0	-0.003	-0.003	0.052	1	1.005	0.005	0.070
	Low -	**Treatment**	0.5	0.499	-0.001	0.046	1	0.998	-0.002	0.065
		Control	-0.5	-0.503	-0.003	0.046	1	1.005	0.005	0.070
	Low +	**Treatment**	-0.5	-0.505	-0.005	0.069	1	1.004	0.004	0.072
		Control	0.5	0.498	-0.002	0.067	1	1.006	0.006	0.070
	High -	**Treatment**	1	0.996	-0.004	0.057	1	1.006	0.006	0.070
		Control	-1	-1.002	-0.002	0.060	1	1.006	0.006	0.074
	High +	**Treatment**	-1	-1.009	-0.009	0.097	1	1.002	0.002	0.075
		Control	1	0.998	-0.002	0.099	1	1.007	0.007	0.077
50 items	**No Difference**	**Treatment**	0	-0.001	-0.001	0.050	1	1.006	0.006	0.062
		Control	0	0.002	0.002	0.050	1	1.005	0.005	0.063
	Low -	**Treatment**	0.5	0.498	-0.002	0.048	1	1.007	0.007	0.062
		Control	-0.5	-0.498	0.002	0.046	1	1.006	0.006	0.062
	Low +	**Treatment**	-0.5	-0.500	0.000	0.064	1	1.007	0.007	0.065
		Control	0.5	0.500	0.000	0.067	1	1.004	0.004	0.067
	High -	**Treatment**	1	0.998	-0.002	0.055	1	1.003	0.003	0.061
		Control	-1	-1.002	-0.002	0.054	1	1.001	0.001	0.063
	High +	**Treatment**	-1	-1.005	-0.005	0.093	1	1.004	0.004	0.070
		Control	1	0.999	-0.001	0.092	1	1.006	0.006	0.071

Table A5

True parameter values and mean IRT estimated parameter values for the outcome in the control and treatment group for the 15 simulations with sample size 10,000, for the ability-based scenario.

			Difficulty				**Discrimination**			
			True	**Est.**	**Bias**	**RMSE**	**True**	**Est.**	**Bias**	**RMSE**
10 items	**No Difference**	**Treatment**	0	-0.002	-0.002	0.038	1	1.004	0.004	0.058
		Control	0	-0.002	-0.002	0.037	1	1.006	0.006	0.059
	Low -	**Treatment**	0.5	0.498	-0.002	0.034	1	1.002	0.002	0.057
		Control	-0.5	-0.500	0.000	0.033	1	1.002	0.002	0.059
	Low +	**Treatment**	-0.5	-0.501	-0.001	0.054	1	1.002	0.002	0.059
		Control	0.5	0.503	0.003	0.055	1	0.999	-0.001	0.062
	High -	**Treatment**	1	1.002	0.002	0.042	1	1.001	0.001	0.056
		Control	-1	-1.002	-0.002	0.043	1	1.002	0.002	0.062
	High +	**Treatment**	-1	-1.003	-0.003	0.080	1	1.001	0.001	0.065
		Control	1	1.003	0.003	0.078	1	1.002	0.002	0.063
20 items	**No Difference**	**Treatment**	0	-0.003	-0.003	0.036	1	1.001	0.001	0.048
		Control	0	-0.003	-0.003	0.036	1	1.000	0.000	0.049
	Low -	**Treatment**	0.5	0.497	-0.003	0.033	1	1.002	0.002	0.047
		Control	-0.5	-0.504	-0.004	0.033	1	1.004	0.004	0.051
	Low +	**Treatment**	-0.5	-0.505	-0.005	0.050	1	1.001	0.001	0.050
		Control	0.5	0.495	-0.005	0.050	1	1.003	0.003	0.052
	High -	**Treatment**	1	0.997	-0.003	0.040	1	1.001	0.001	0.048
		Control	-1	-1.005	-0.005	0.041	1	1.001	0.001	0.051
	High +	**Treatment**	-1	-1.000	0.000	0.067	1	1.005	0.005	0.054
		Control	1	0.997	-0.003	0.071	1	1.003	0.003	0.056
50 items	**No Difference**	**Treatment**	0	0.000	0.000	0.035	1	1.002	0.002	0.043
		Control	0	0.003	0.003	0.036	1	1.003	0.003	0.044
	Low -	**Treatment**	0.5	0.499	-0.001	0.034	1	1.004	0.004	0.045
		Control	-0.5	-0.498	0.002	0.034	1	1.004	0.004	0.042
	Low +	**Treatment**	-0.5	-0.499	0.001	0.046	1	1.004	0.004	0.045
		Control	0.5	0.500	0.000	0.046	1	1.005	0.005	0.048
	High -	**Treatment**	1	0.996	-0.004	0.039	1	1.005	0.005	0.044
		Control	-1	-0.999	0.001	0.041	1	1.005	0.005	0.047
	High +	**Treatment**	-1	-0.998	0.002	0.066	1	1.007	0.007	0.049
		Control	1	1.001	0.001	0.065	1	1.002	0.002	0.050

Table A6

True parameter values and mean IRT estimated parameter values for the outcome in the control and treatment group for the 15 simulations with sample size of 500 for the cut-score scenario.

			Difficulty				**Discrimination**			
			True	**Est.**	**Bias**	**RMSE**	**True**	**Est.**	**Bias**	**RMSE**
10 items	**No Difference**	**Treatment**	0	0.006	0.006	0.140	1	1.023	0.023	0.267
		Control	0	-13.474	-13.474	51.106	1	1.811	0.811	6.426
	Low -	**Treatment**	0.5	0.528	0.028	0.234	1	1.042	0.042	0.289
		Control	-0.5	-21.194	-20.694	66.266	1	1.837	0.837	5.386
	Low +	**Treatment**	-0.5	-0.513	-0.013	0.129	1	1.026	0.026	0.257
		Control	0.5	-9.852	-10.352	41.152	1	1.245	0.245	1.756
	High -	**Treatment**	1	1.081	0.081	0.425	1	1.029	0.029	0.301
		Control	-1	-35.313	-34.313	91.613	1	6.083	5.083	22.522
	High +	**Treatment**	-1	-1.034	-0.034	0.210	1	1.014	0.014	0.259
		Control	1	-1.168	-2.168	23.888	1	1.419	0.419	2.534
20 items	**No Difference**	**Treatment**	0	0.012	0.012	0.142	1	1.004	0.004	0.212
		Control	0	-8.315	-8.315	39.022	1	1.178	0.178	0.842
	Low -	**Treatment**	0.5	0.514	0.014	0.204	1	1.031	0.031	0.233
		Control	-0.5	-19.082	-18.582	63.397	1	1.323	0.323	3.217
	Low +	**Treatment**	-0.5	-0.507	-0.007	0.133	1	1.009	0.009	0.209
		Control	0.5	-6.595	-7.095	35.545	1	1.132	0.132	0.737
	High -	**Treatment**	1	1.052	0.052	0.351	1	1.029	0.029	0.255
		Control	-1	-32.730	-31.730	90.238	1	2.624	1.624	10.540
	High +	**Treatment**	-1	-1.024	-0.024	0.188	1	1.011	0.011	0.204
		Control	1	-0.776	-1.776	16.857	1	1.132	0.132	0.702
50 items	**No Difference**	**Treatment**	0	0.010	0.010	0.147	1	1.022	0.022	0.210
		Control	0	-3.271	-3.271	22.743	1	1.064	0.064	0.570
	Low -	**Treatment**	0.5	0.532	0.032	0.242	1	1.022	0.022	0.227
		Control	-0.5	-7.126	-6.626	38.215	1	1.174	0.174	0.710
	Low +	**Treatment**	-0.5	-0.504	-0.004	0.131	1	1.017	0.017	0.197
		Control	0.5	-1.293	-1.793	16.703	1	1.057	0.057	0.514
	High -	**Treatment**	1	1.058	0.058	0.371	1	1.023	0.023	0.251
		Control	-1	-18.154	-17.154	64.696	1	1.155	0.155	0.842
	High +	**Treatment**	-1	-1.015	-0.015	0.176	1	1.017	0.017	0.201
		Control	1	0.709	-0.291	5.086	1	1.035	0.035	0.464

Table A7

True parameter values and mean IRT estimated parameter values for the outcome in the control and treatment group for the 15 simulations with sample size of 1,000 for the cut-score scenario.

			Difficulty				Discrimination			
			True	**Est.**	**Bias**	**RMSE**	**True**	**Est.**	**Bias**	**RMSE**
10 items	**No Difference**	**Treatment**	0	0.005	0.005	0.093	1	1.017	0.017	0.190
		Control	0	-4.717	-4.717	27.503	1	1.234	0.234	3.305
	Low -	**Treatment**	0.5	0.518	0.018	0.159	1	1.015	0.015	0.190
		Control	-0.5	-9.055	-8.555	40.689	1	1.128	0.128	0.798
	Low +	**Treatment**	-0.5	-0.505	-0.005	0.092	1	1.004	0.004	0.179
		Control	0.5	-1.582	-2.082	16.261	1	1.079	0.079	0.713
	High -	**Treatment**	1	1.033	0.033	0.244	1	1.012	0.012	0.205
		Control	-1	-14.545	-13.545	53.699	1	1.527	0.527	5.003
	High +	**Treatment**	-1	-1.017	-0.017	0.149	1	1.006	0.006	0.183
		Control	1	0.462	-0.538	10.094	1	1.068	0.068	0.579
20 items	**No Difference**	**Treatment**	0	0.000	0.000	0.097	1	1.012	0.012	0.157
		Control	0	-2.946	-2.946	21.258	1	1.067	0.067	0.533
	Low -	**Treatment**	0.5	0.517	0.017	0.146	1	1.004	0.004	0.161
		Control	-0.5	-4.424	-3.924	27.080	1	1.126	0.126	0.670
	Low +	**Treatment**	-0.5	-0.505	-0.005	0.089	1	0.999	-0.001	0.143
		Control	0.5	-0.232	-0.732	9.679	1	1.055	0.055	0.483
	High -	**Treatment**	1	1.021	0.021	0.227	1	1.017	0.017	0.181
		Control	-1	-12.669	-11.669	50.164	1	1.242	0.242	1.412
	High +	**Treatment**	-1	-1.018	-0.018	0.122	1	1.000	0.000	0.144
		Control	1	0.728	-0.272	4.412	1	1.049	0.049	0.438
50 items	**No Difference**	**Treatment**	0	0.008	0.008	0.099	1	1.006	0.006	0.142
		Control	0	-0.584	-0.584	7.766	1	1.046	0.046	0.379
	Low -	**Treatment**	0.5	0.511	0.011	0.163	1	1.011	0.011	0.166
		Control	-0.5	-1.765	-1.265	12.661	1	1.033	0.033	0.455
	Low +	**Treatment**	-0.5	-0.501	-0.001	0.089	1	1.011	0.011	0.135
		Control	0.5	0.428	-0.072	0.640	1	1.070	0.070	0.347
	High -	**Treatment**	1	1.020	0.020	0.257	1	1.015	0.015	0.179
		Control	-1	-5.020	-4.020	27.623	1	1.111	0.111	0.577
	High +	**Treatment**	-1	-1.006	-0.006	0.110	1	1.016	0.016	0.134
		Control	1	0.989	-0.011	0.161	1	1.022	0.022	0.313

Table A8

True parameter values and mean IRT estimated parameter values for the outcome in the control and treatment group for the 15 simulations with sample size of 5,000 for the cut-score scenario.

			Difficulty				Discrimination			
			True	**Est.**	**Bias**	**RMSE**	**True**	**Est.**	**Bias**	**RMSE**
10 items	**No Difference**	**Treatment**	0	0.000	0.000	0.043	1	0.998	-0.002	0.078
		Control	0	-0.040	-0.040	0.298	1	1.033	0.033	0.273
	Low -	**Treatment**	0.5	0.504	0.004	0.064	1	1.000	0.000	0.082
		Control	-0.5	-0.593	-0.093	0.472	1	1.023	0.023	0.282
	Low +	**Treatment**	-0.5	-0.501	-0.001	0.038	1	1.006	0.006	0.080
		Control	0.5	0.474	-0.026	0.158	1	1.012	0.012	0.239
	High -	**Treatment**	1	1.007	0.007	0.106	1	1.001	0.001	0.091
		Control	-1	-1.167	-0.167	0.830	1	1.044	0.044	0.354
	High +	**Treatment**	-1	-1.003	-0.003	0.058	1	1.004	0.004	0.079
		Control	1	0.999	-0.001	0.074	1	1.023	0.023	0.229
20 items	**No Difference**	**Treatment**	0	-0.002	-0.002	0.041	1	1.004	0.004	0.066
		Control	0	-0.059	-0.059	0.302	1	1.008	0.008	0.227
	Low -	**Treatment**	0.5	0.501	0.001	0.058	1	1.001	0.001	0.068
		Control	-0.5	-0.597	-0.097	0.549	1	1.020	0.020	0.261
	Low +	**Treatment**	-0.5	-0.505	-0.005	0.040	1	1.001	0.001	0.063
		Control	0.5	0.482	-0.018	0.166	1	1.020	0.020	0.198
	High -	**Treatment**	1	1.001	0.001	0.094	1	1.005	0.005	0.079
		Control	-1	-1.238	-0.238	1.535	1	1.018	0.018	0.306
	High +	**Treatment**	-1	-1.004	-0.004	0.052	1	1.001	0.001	0.062
		Control	1	0.988	-0.012	0.083	1	1.009	0.009	0.180
50 items	**No Difference**	**Treatment**	0	0.001	0.001	0.042	1	1.003	0.003	0.064
		Control	0	-0.034	-0.034	0.209	1	1.008	0.008	0.173
	Low -	**Treatment**	0.5	0.503	0.003	0.064	1	1.003	0.003	0.068
		Control	-0.5	-0.558	-0.058	0.372	1	1.009	0.009	0.200
	Low +	**Treatment**	-0.5	-0.500	0.000	0.041	1	1.004	0.004	0.063
		Control	0.5	0.489	-0.011	0.109	1	1.016	0.016	0.155
	High -	**Treatment**	1	1.008	0.008	0.107	1	1.002	0.002	0.080
		Control	-1	-1.113	-0.113	0.619	1	1.015	0.015	0.245
	High +	**Treatment**	-1	-1.004	-0.004	0.051	1	1.003	0.003	0.059
		Control	1	0.989	-0.011	0.067	1	1.005	0.005	0.134

Table A9

True parameter values and mean IRT estimated parameter values for the outcome in the control and treatment group for the 15 simulations with sample size of 10,000 for the cut-score scenario.

			Difficulty				**Discrimination**			
			True	**Est.**	**Bias**	**RMSE**	**True**	**Est.**	**Bias**	**RMSE**
10 items	**No Difference**	**Treatment**	0	0.000	0.000	0.031	1	1.002	0.002	0.056
		Control	0	-0.022	-0.022	0.183	1	1.011	0.011	0.175
	Low -	**Treatment**	0.5	0.499	-0.001	0.044	1	1.006	0.006	0.057
		Control	-0.5	-0.558	-0.058	0.301	1	0.999	-0.001	0.193
	Low +	**Treatment**	-0.5	-0.500	0.000	0.028	1	1.000	0.000	0.058
		Control	0.5	0.491	-0.009	0.103	1	1.014	0.014	0.165
	High -	**Treatment**	1	1.006	0.006	0.069	1	0.998	-0.002	0.061
		Control	-1	-1.077	-0.077	0.473	1	1.015	0.015	0.230
	High +	**Treatment**	-1	-1.001	-0.001	0.042	1	1.002	0.002	0.057
		Control	1	1.002	0.002	0.052	1	1.021	0.021	0.159
20 items	**No Difference**	**Treatment**	0	-0.002	-0.002	0.028	1	1.002	0.002	0.045
		Control	0	-0.030	-0.030	0.200	1	1.004	0.004	0.156
	Low -	**Treatment**	0.5	0.501	0.001	0.042	1	1.000	0.000	0.049
		Control	-0.5	-0.550	-0.050	0.323	1	1.007	0.007	0.183
	Low +	**Treatment**	-0.5	-0.505	-0.005	0.028	1	1.004	0.004	0.045
		Control	0.5	0.474	-0.026	0.113	1	0.990	-0.010	0.141
	High -	**Treatment**	1	1.000	0.000	0.066	1	1.001	0.001	0.056
		Control	-1	-1.066	-0.066	0.479	1	1.018	0.018	0.214
	High +	**Treatment**	-1	-1.003	-0.003	0.036	1	1.001	0.001	0.044
		Control	1	0.991	-0.009	0.058	1	1.000	0.000	0.134
50 items	**No Difference**	**Treatment**	0	-0.001	-0.001	0.031	1	1.007	0.007	0.047
		Control	0	-0.015	-0.015	0.138	1	1.007	0.007	0.118
	Low -	**Treatment**	0.5	0.501	0.001	0.048	1	1.002	0.002	0.050
		Control	-0.5	-0.527	-0.027	0.220	1	1.004	0.004	0.131
	Low +	**Treatment**	-0.5	-0.501	-0.001	0.028	1	1.002	0.002	0.043
		Control	0.5	0.489	-0.011	0.074	1	1.004	0.004	0.105
	High -	**Treatment**	1	1.001	0.001	0.069	1	1.004	0.004	0.054
		Control	-1	-1.027	-0.027	0.348	1	1.017	0.017	0.162
	High +	**Treatment**	-1	-1.004	-0.004	0.036	1	1.002	0.002	0.043
		Control	1	0.992	-0.008	0.046	1	1.002	0.002	0.094

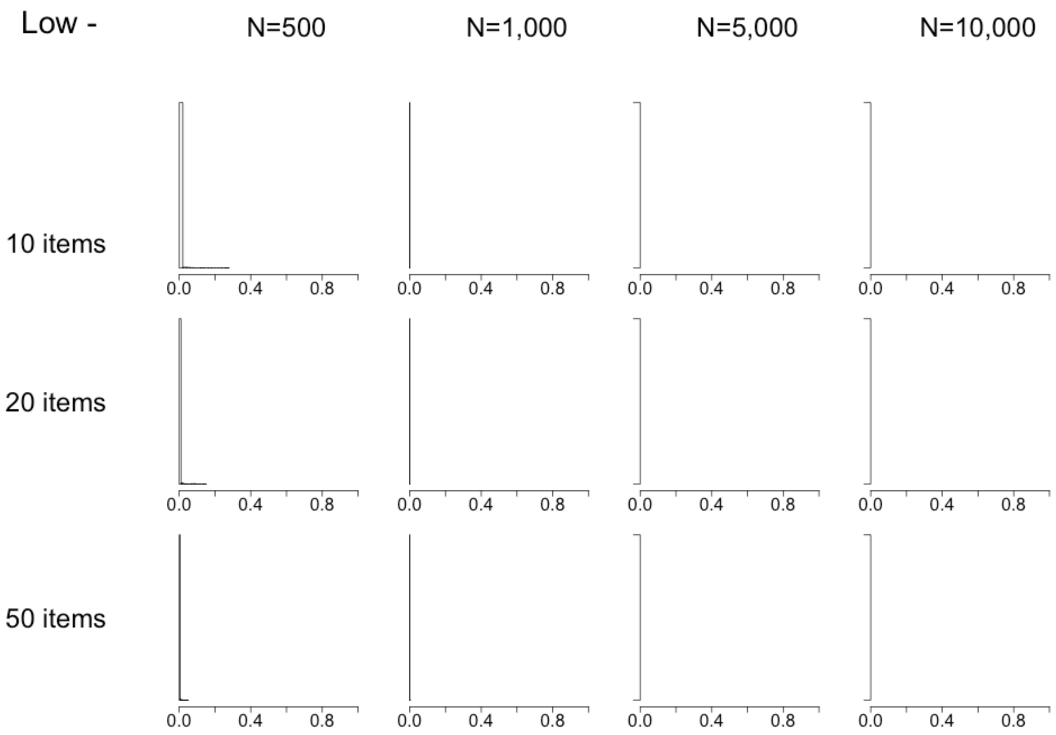

Figure A1: Histograms of the *p*-values from the IRT method under the ability-based scenario for the Low - outcome condition.

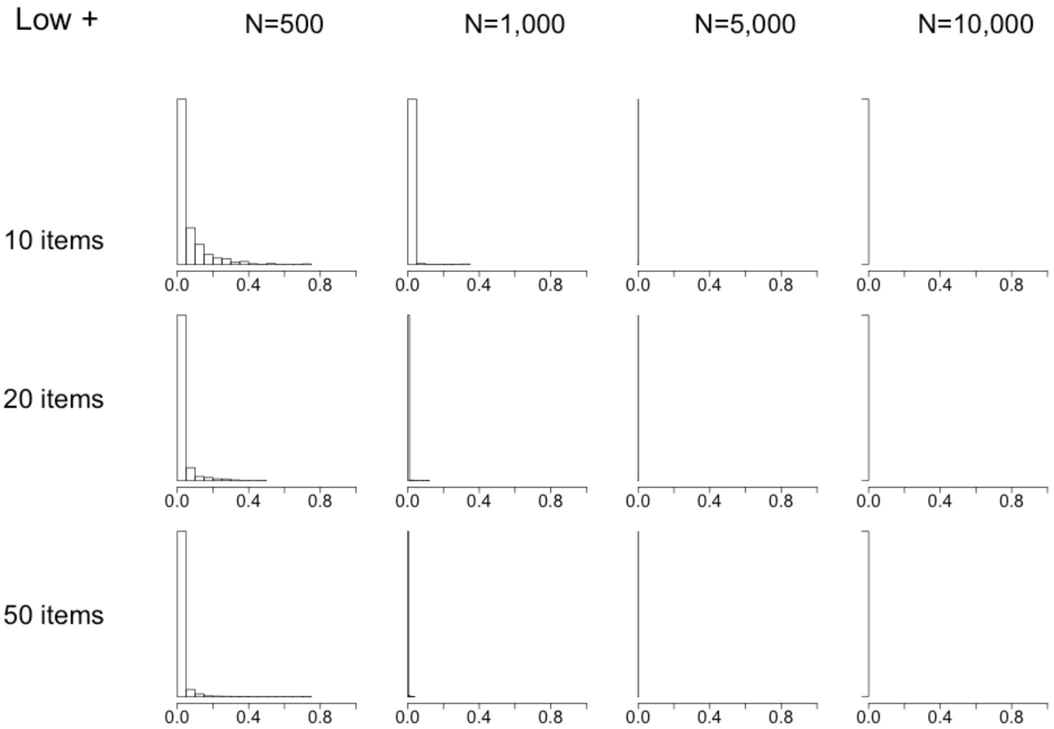

Figure A2: Histograms of the *p*-values from the IRT method under the ability-based scenario for the Low + outcome condition.

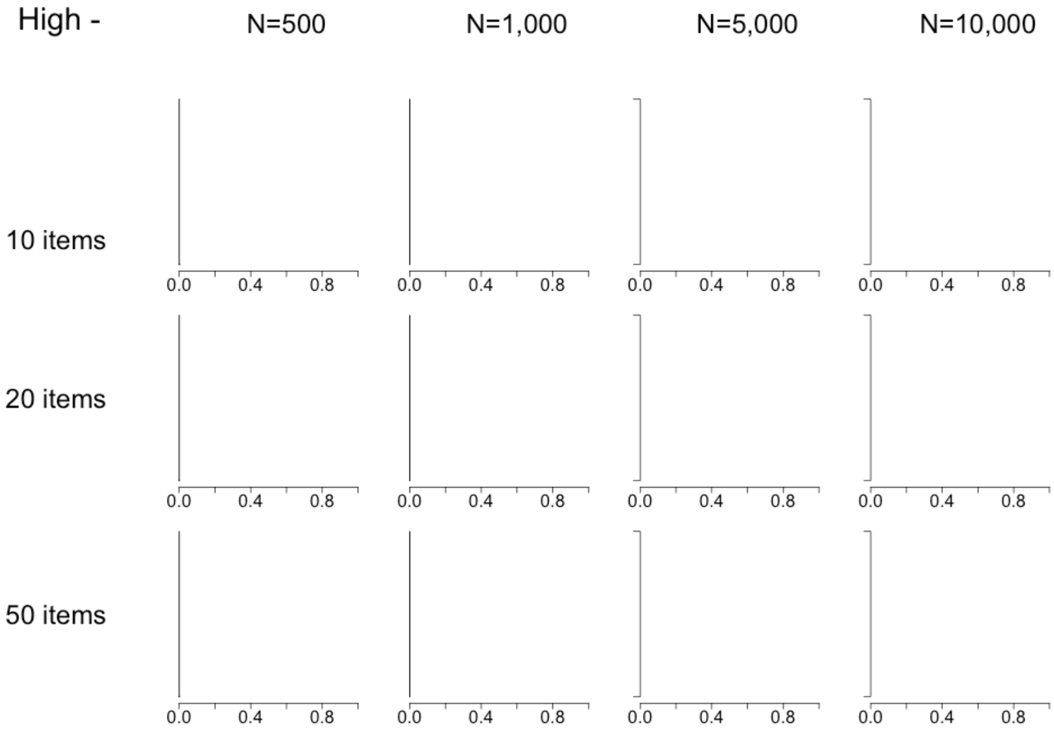

Figure A3: Histograms of the *p*-values from the IRT method under the ability-based scenario for the High - outcome condition.

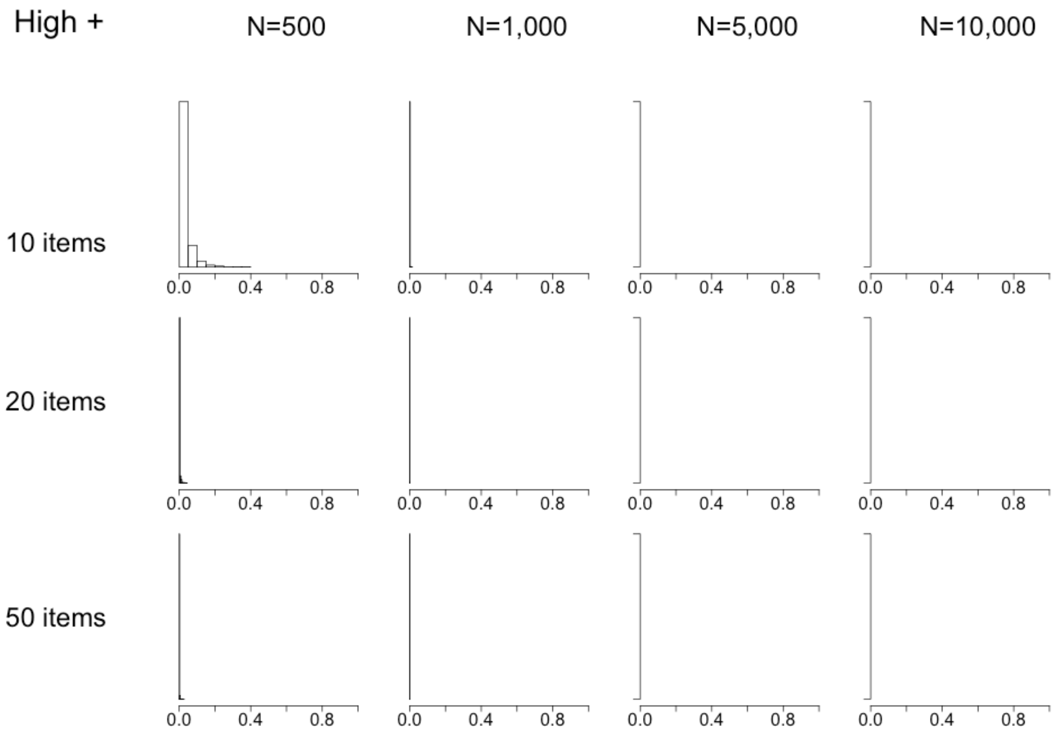

Figure A4: Histograms of the *p*-values from the IRT method under the ability-based scenario for the High + outcome condition.

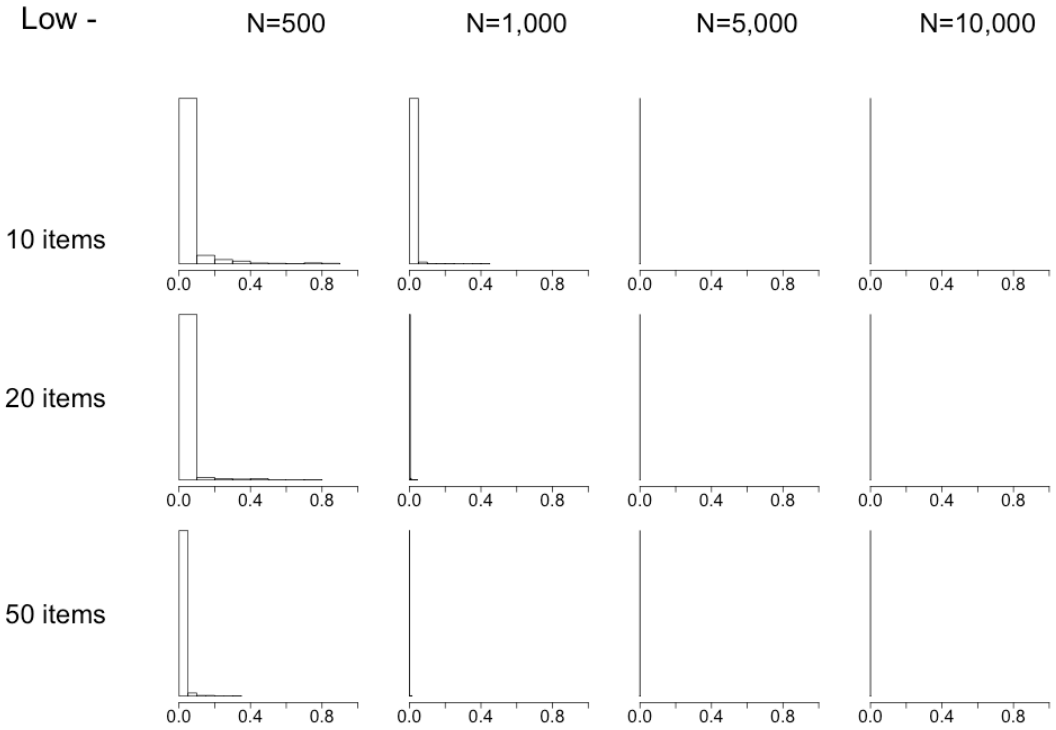

Figure A5: Histograms of the *p*-values from the LR method under the ability-based scenario for the Low - outcome condition.

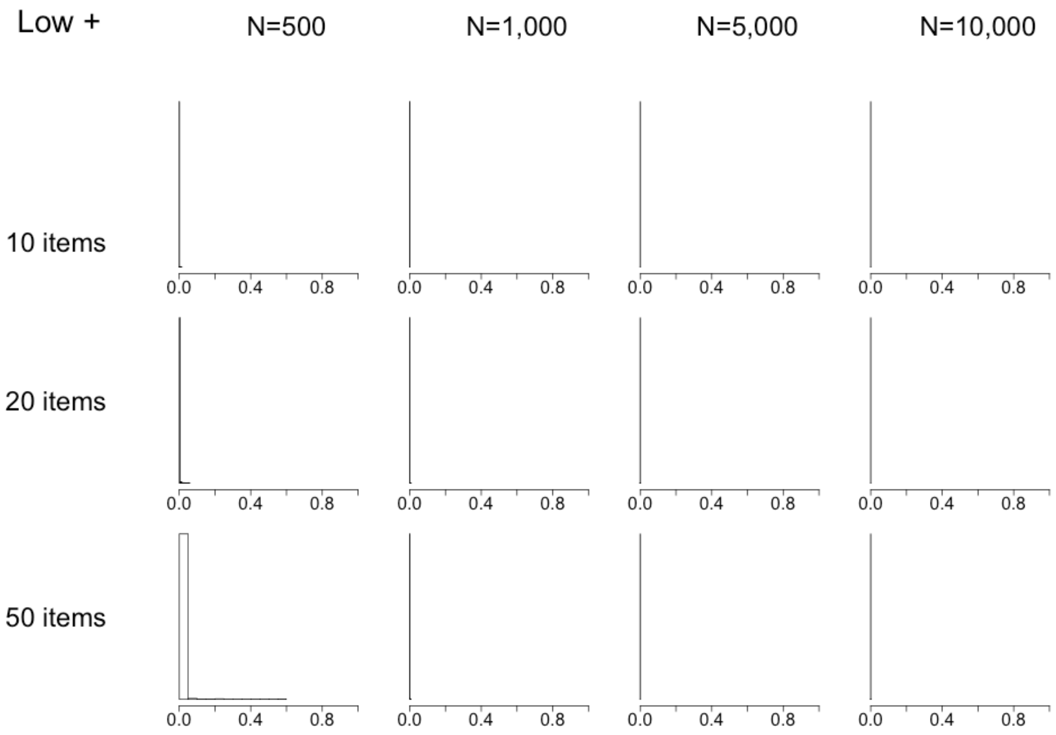

Figure A6: Histograms of the *p*-values from the LR method under the ability-based scenario for the Low + outcome condition.

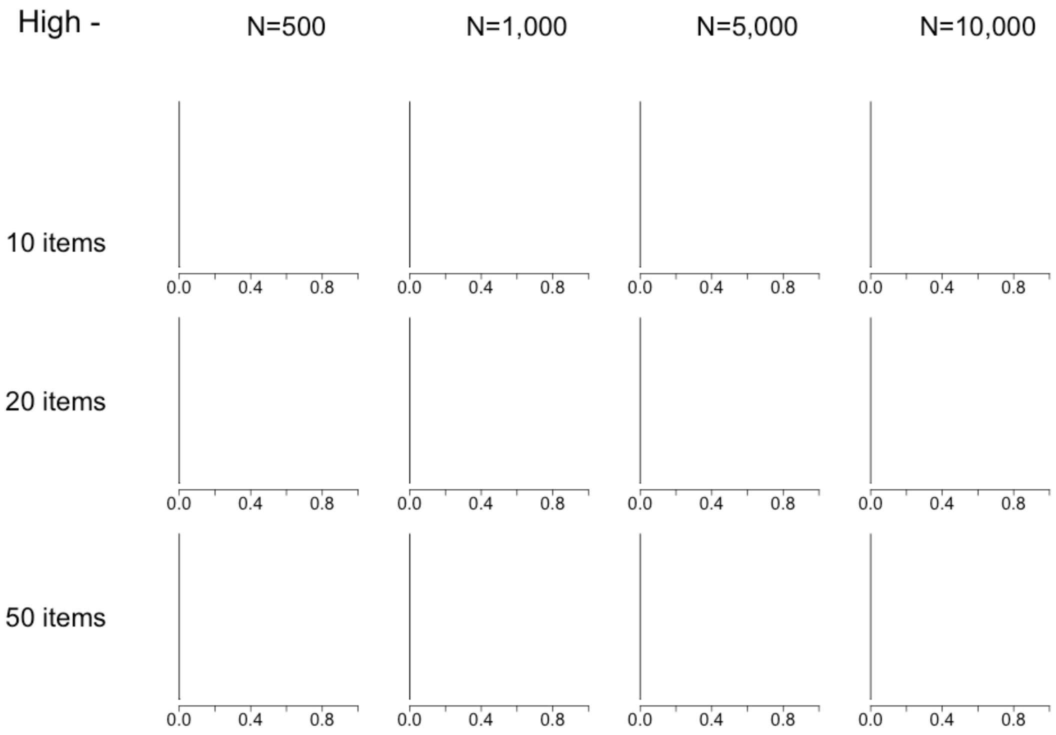

Figure A7: Histograms of the *p*-values from the LR method under the ability-based scenario for the High - outcome condition.

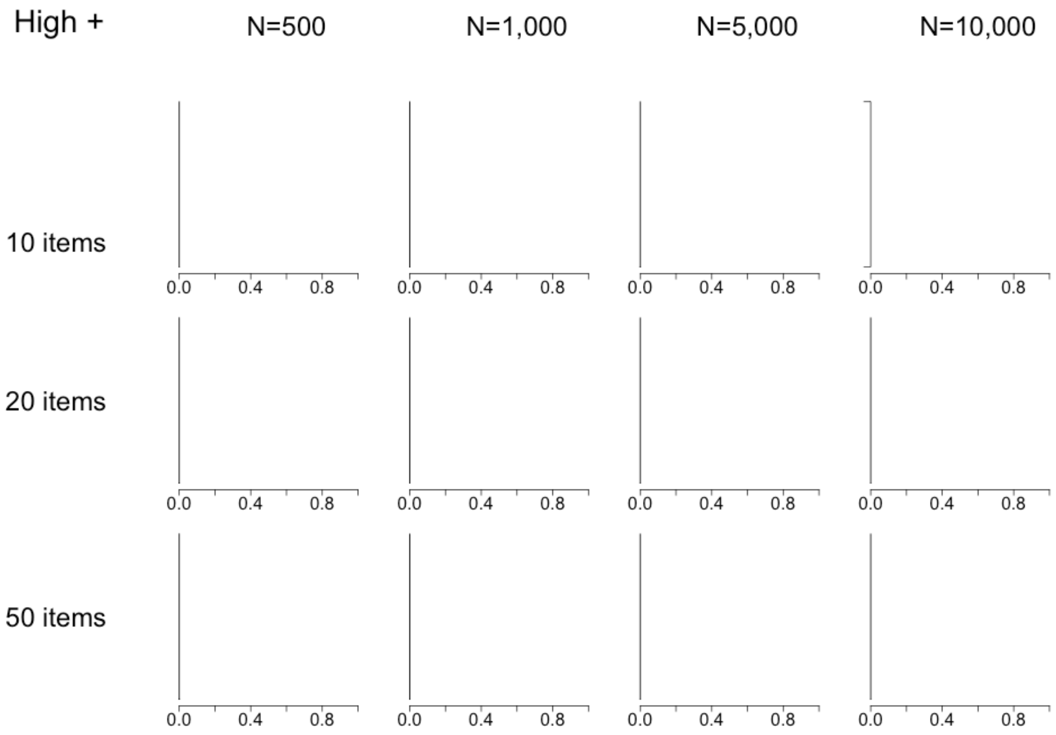

Figure A8: Histograms of the *p*-values from the LR method under the ability-based scenario for the High + outcome condition.

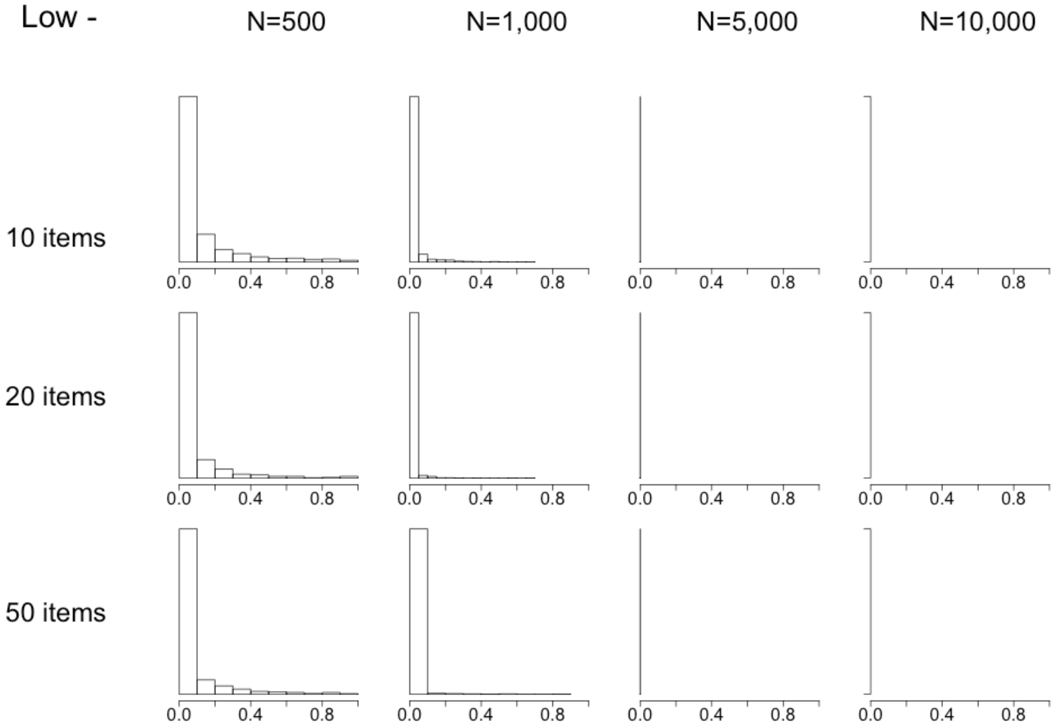

Figure A9: Histograms of the *p*-values from the propensity score method under the ability-based scenario for the Low - outcome condition.

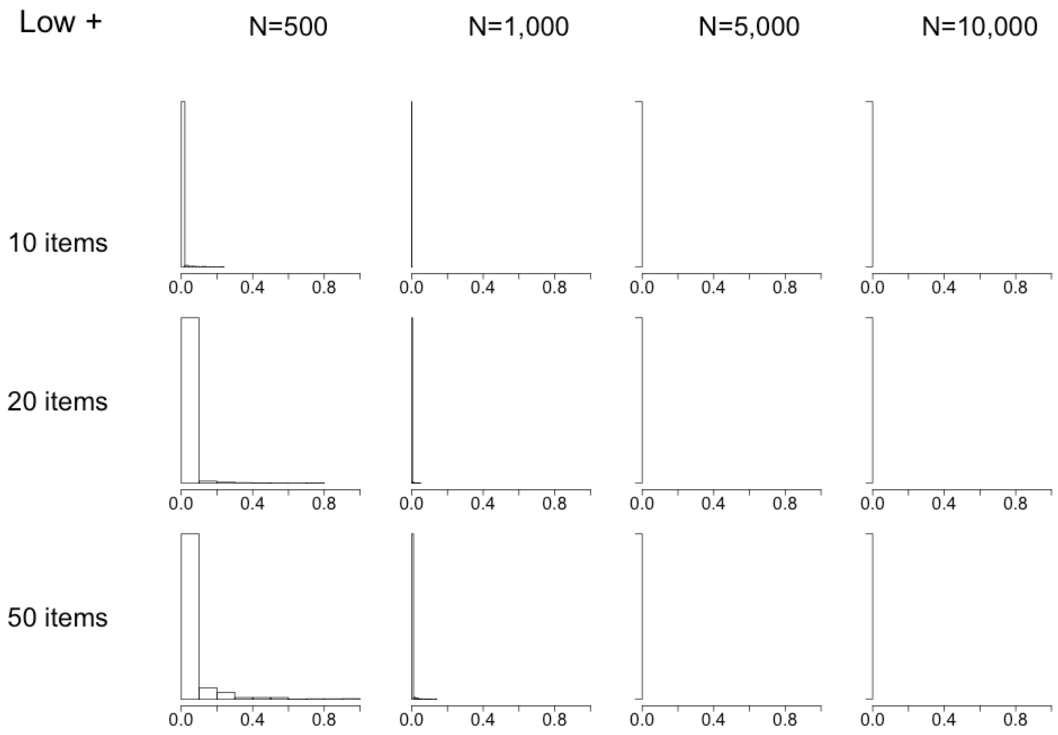

Figure A10: Histograms of the *p*-values from the propensity score method under the ability-based scenario for the Low + outcome condition.

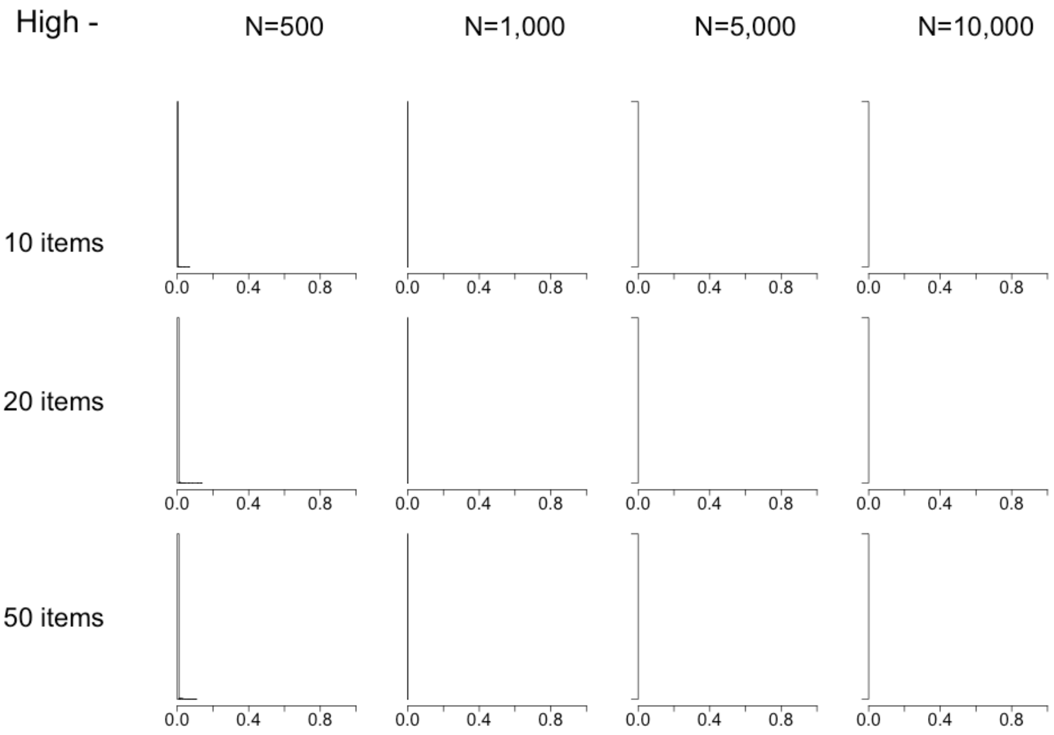

Figure A11: Histograms of the *p*-values from the propensity score method under the ability-based scenario for the High - outcome condition.

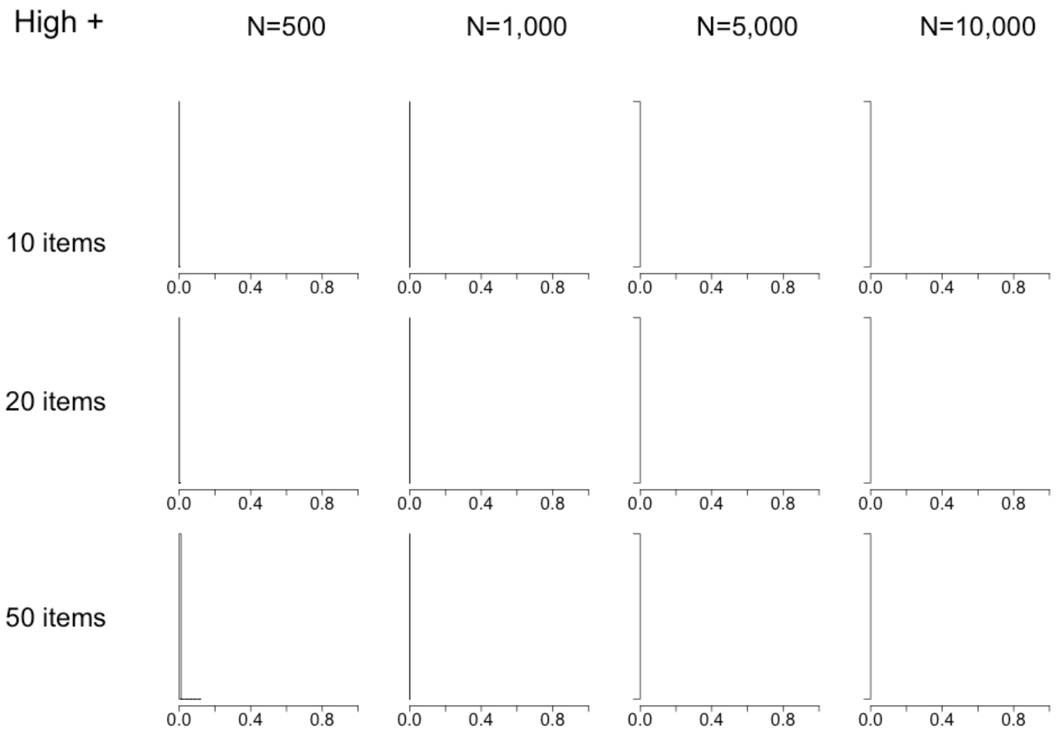

Figure A12: Histograms of the *p*-values from the propensity score method under the ability-based scenario for the High + outcome condition.

.

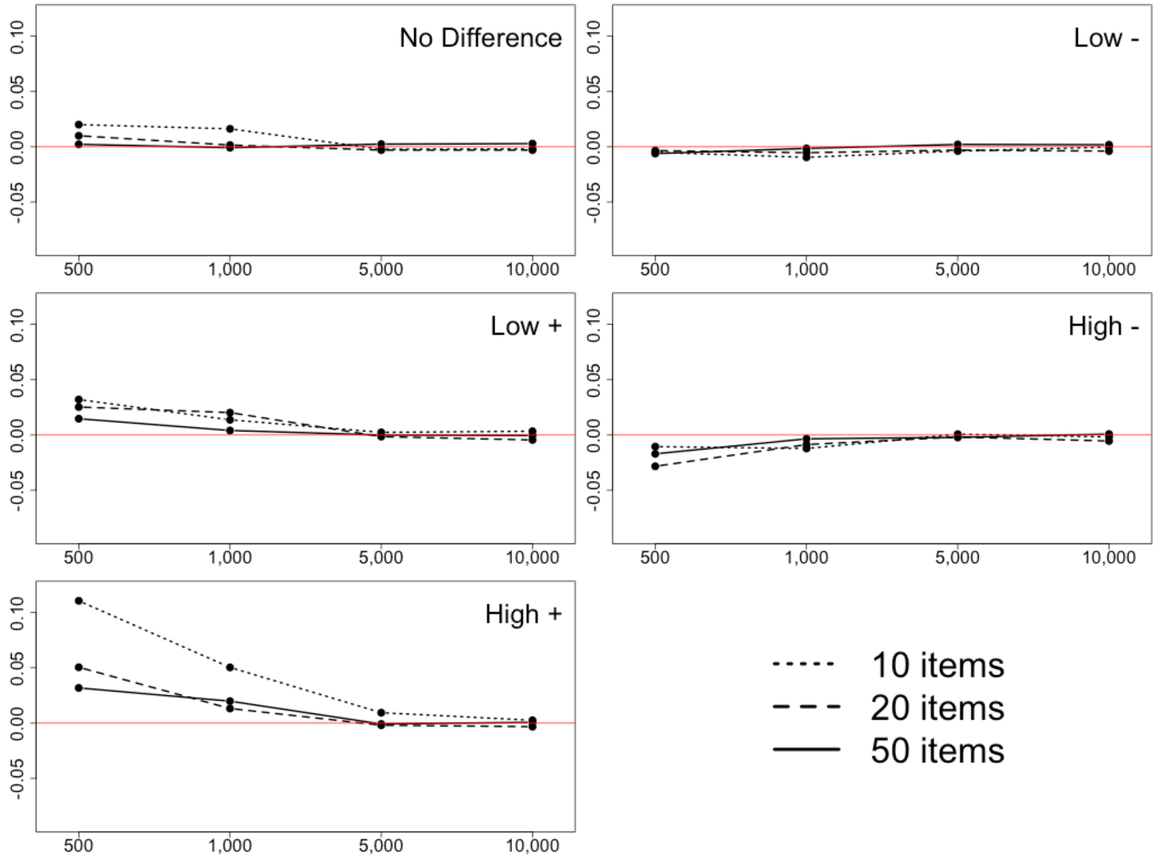

Figure A13: Profile plot for the bias of the difficulty parameter of the outcome variable for the control group under the ability-based scenario.

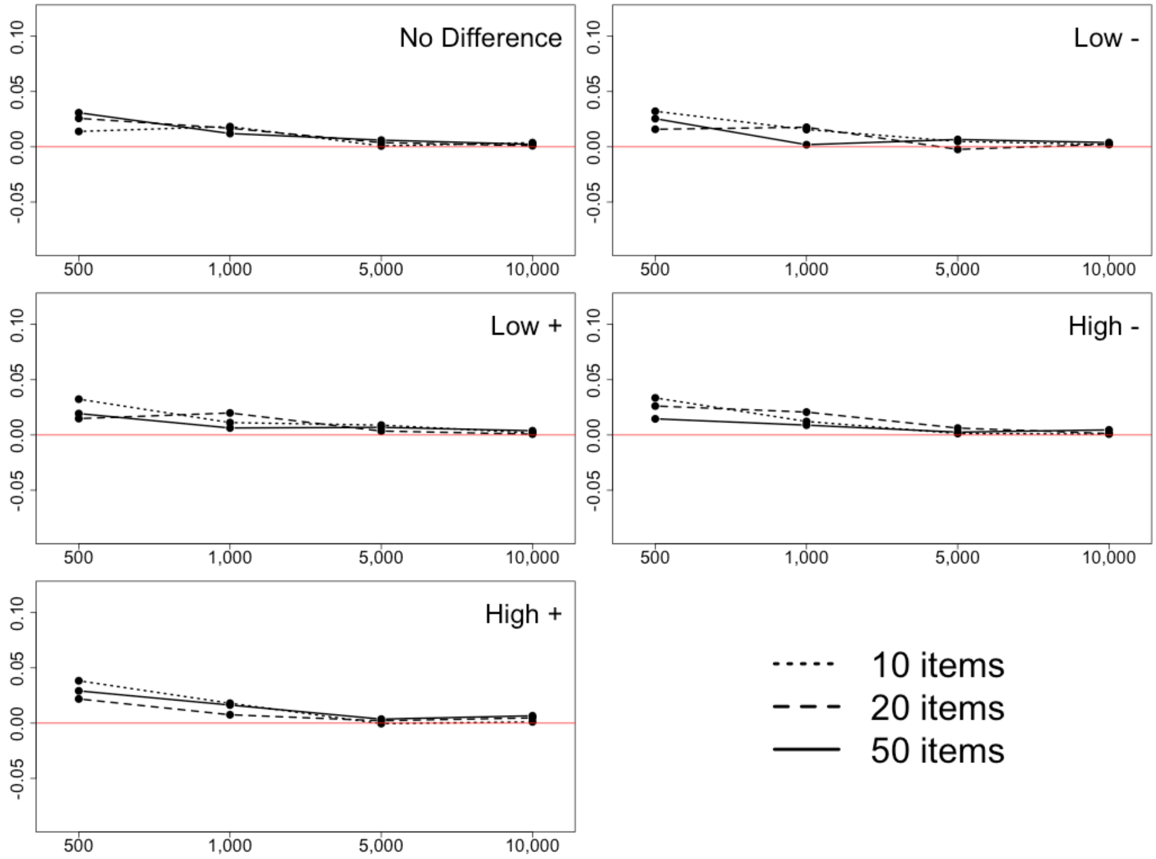

Figure A14: Profile plot for the bias of the discrimination parameter of the outcome variable for the treatment group under the ability-based scenario.

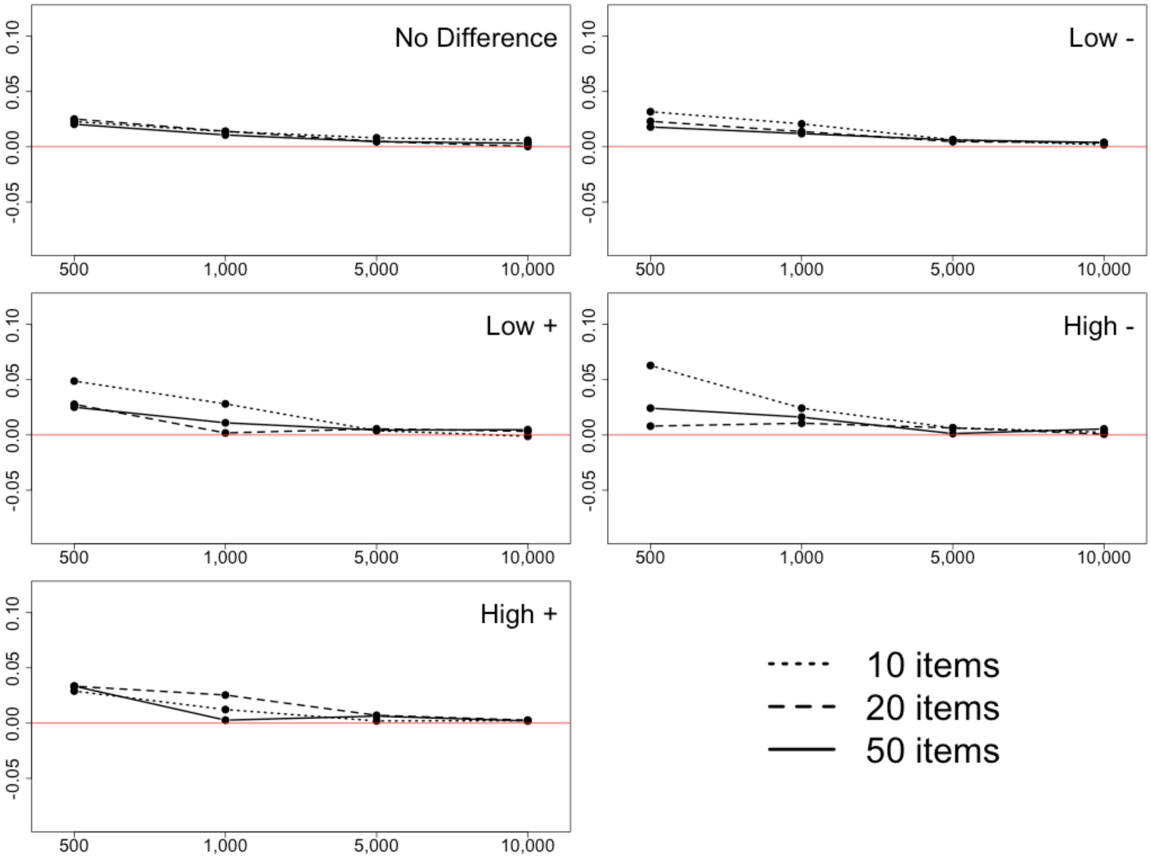

Figure A15: Profile plot for the bias of the discrimination parameter of the outcome variable for the control group under the ability-based scenario.

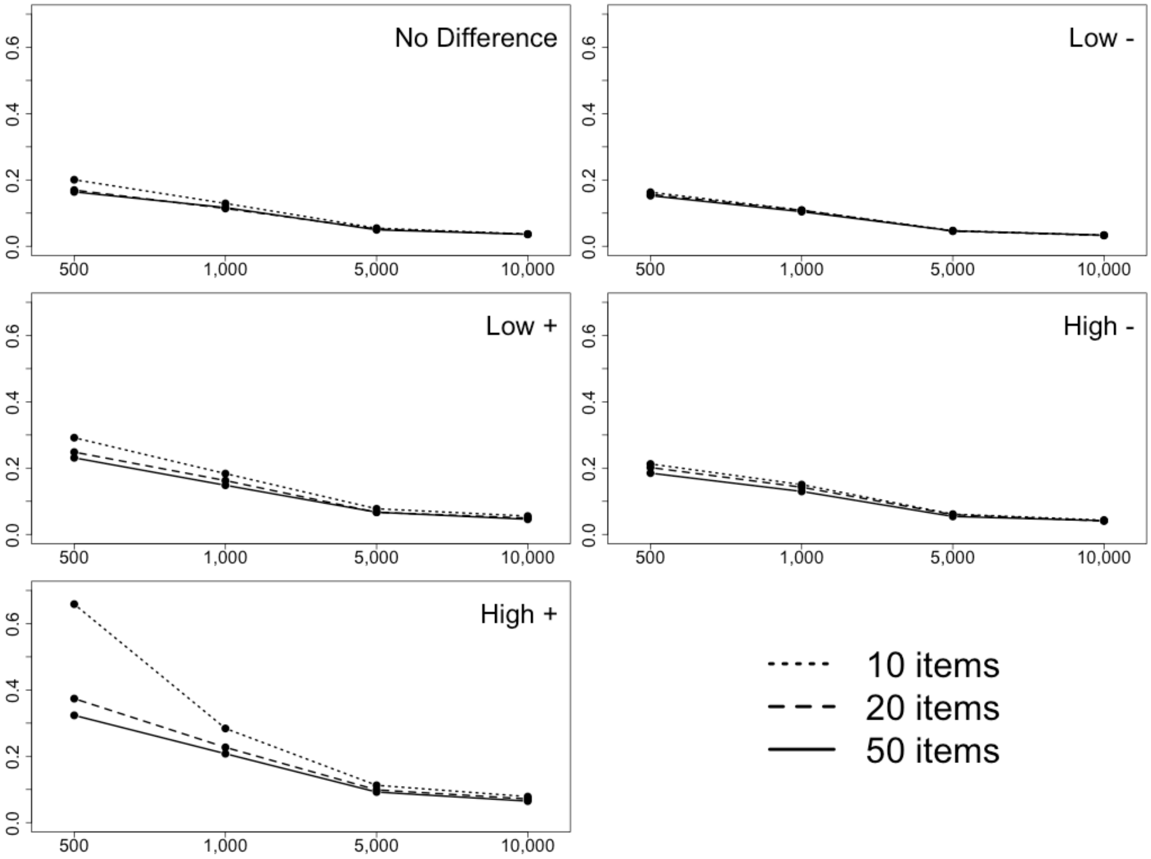

Figure A16: Profile plot for the RMSE of the difficulty parameter of the outcome variable for the control group under the ability-based scenario.

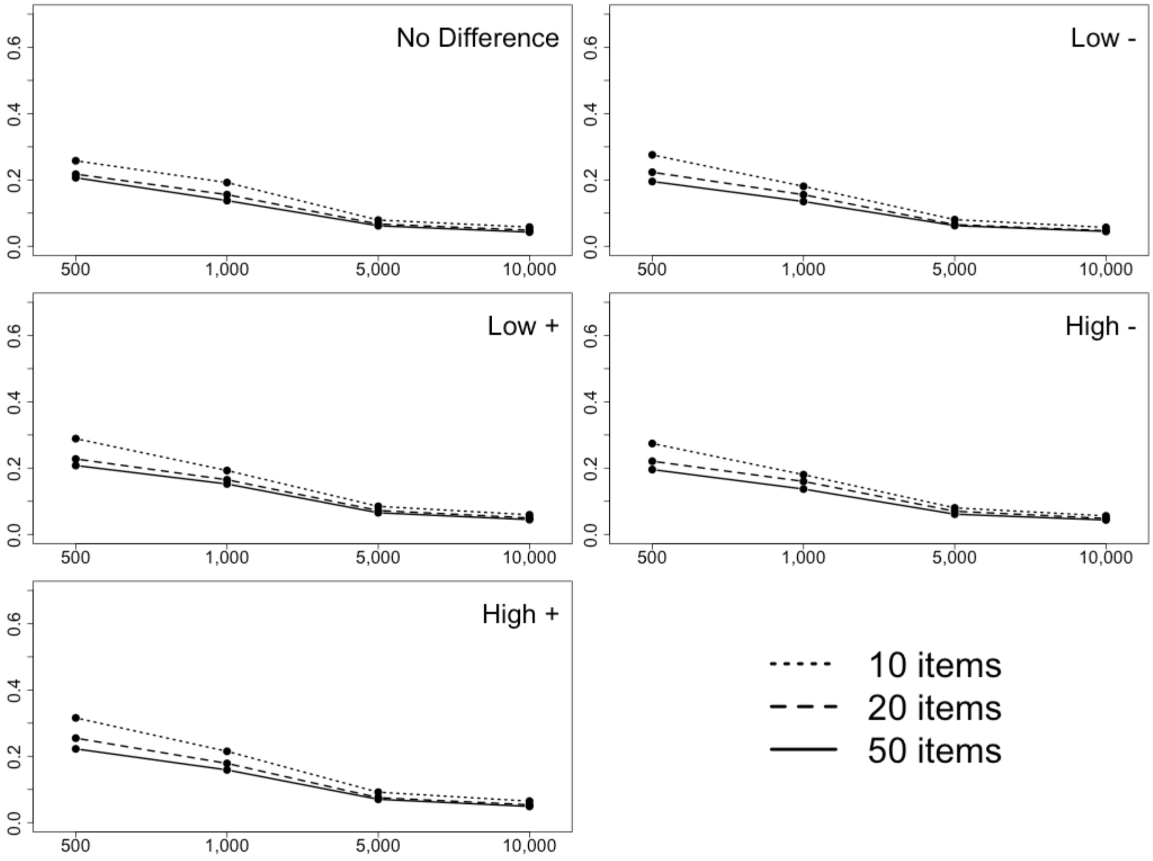

Figure A17: Profile plot for the RMSE of the discrimination parameter of the outcome variable for the treatment group under the ability-based scenario.

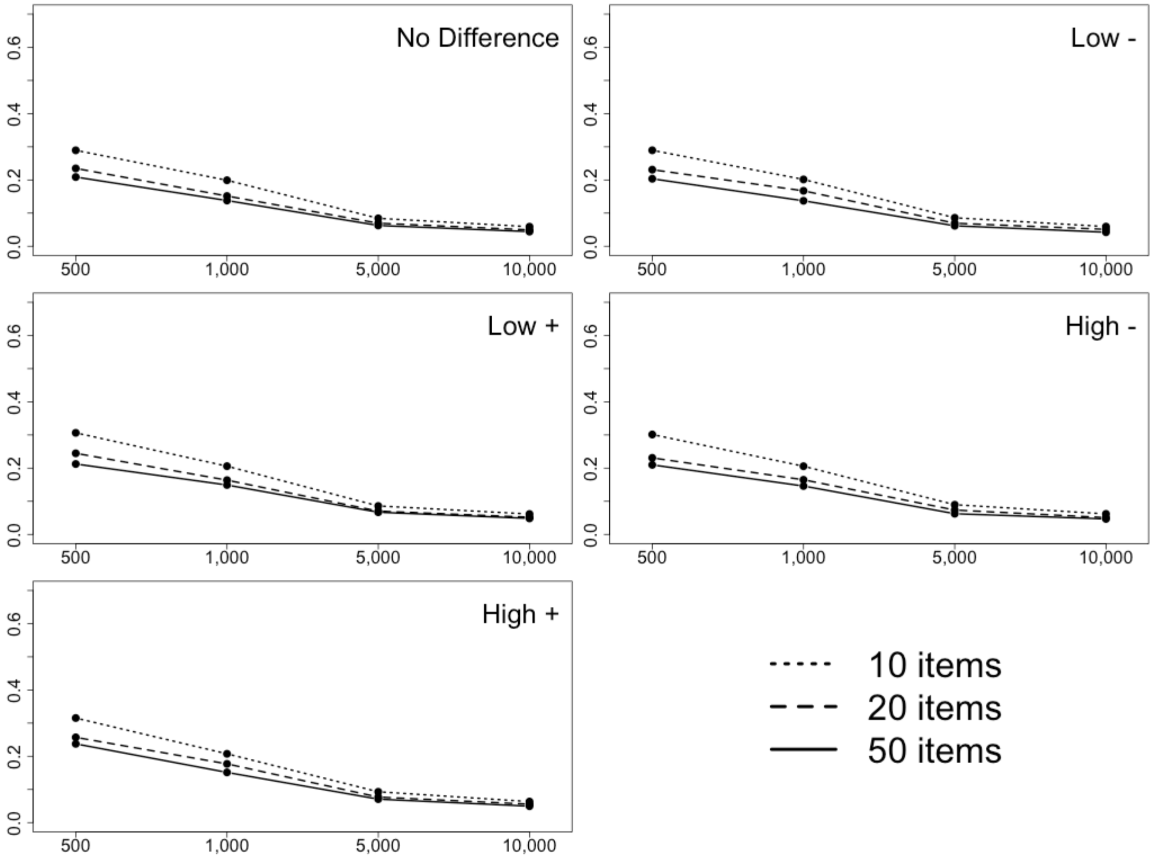

Figure A18: Profile plot for the RMSE of the discrimination parameter of the outcome variable for the control group under the ability-based scenario.

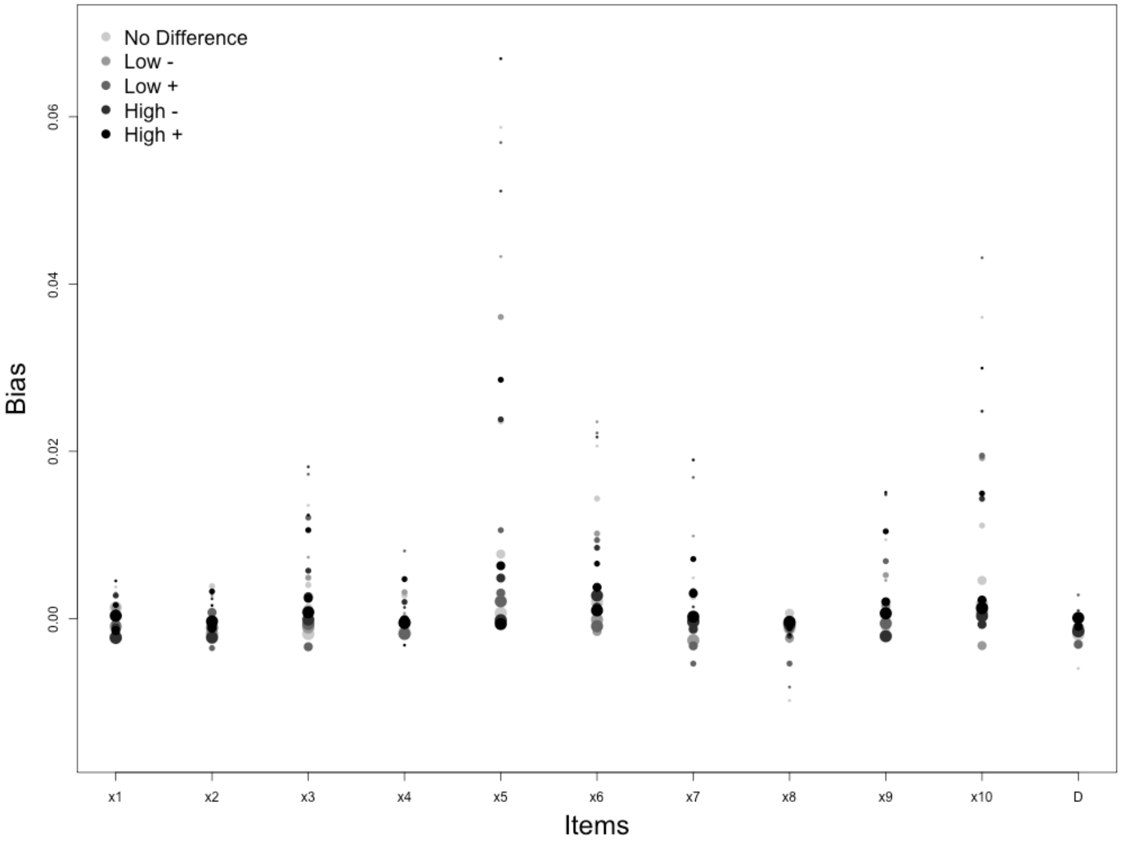

Figure A19: Plot of bias difficulty for the 10-item test under the ability-based scenario.

Note: The sample size is represented by the increase size in dots on the plot with the smallest dot representing N = 500 and the largest representing N = 10,000.

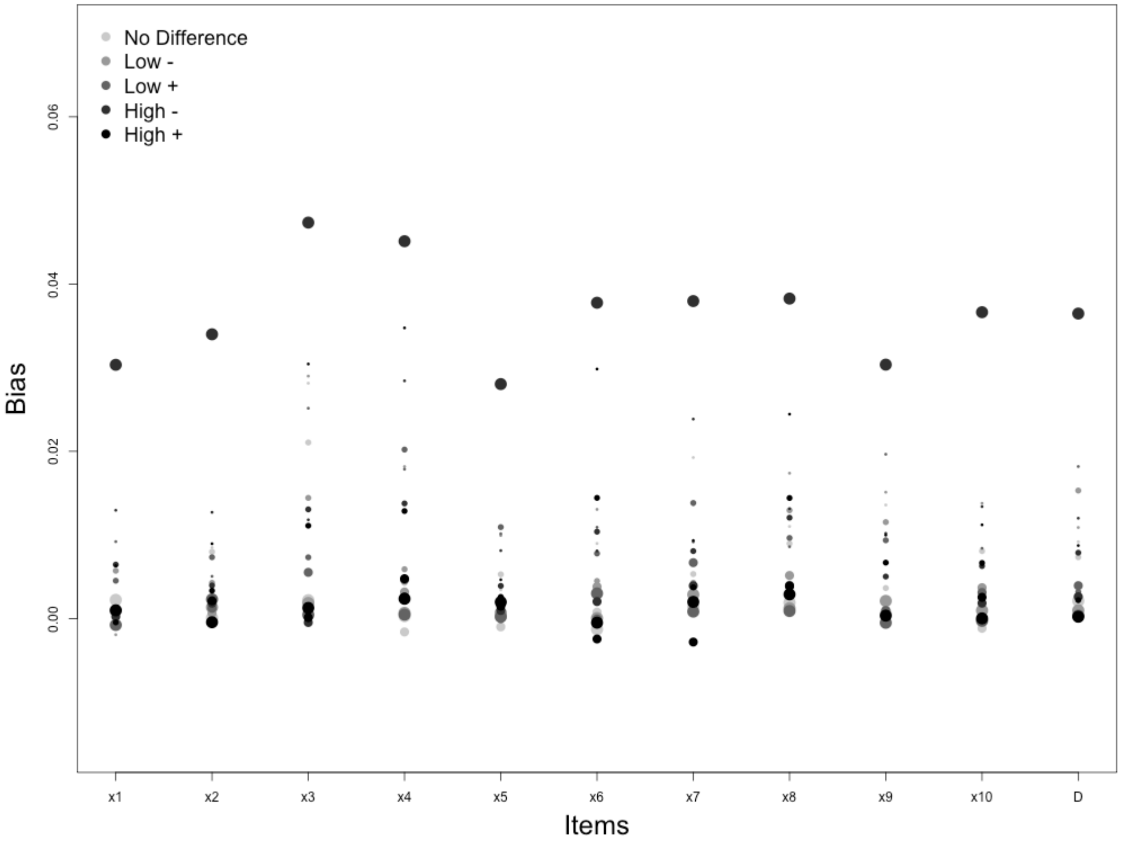

Figure A20: Plot of bias discrimination for the 10-item test under the ability-based scenario.

Note: The sample size is represented by the increase size in dots on the plot with the smallest dot representing N = 500 and the largest representing N = 10,000.

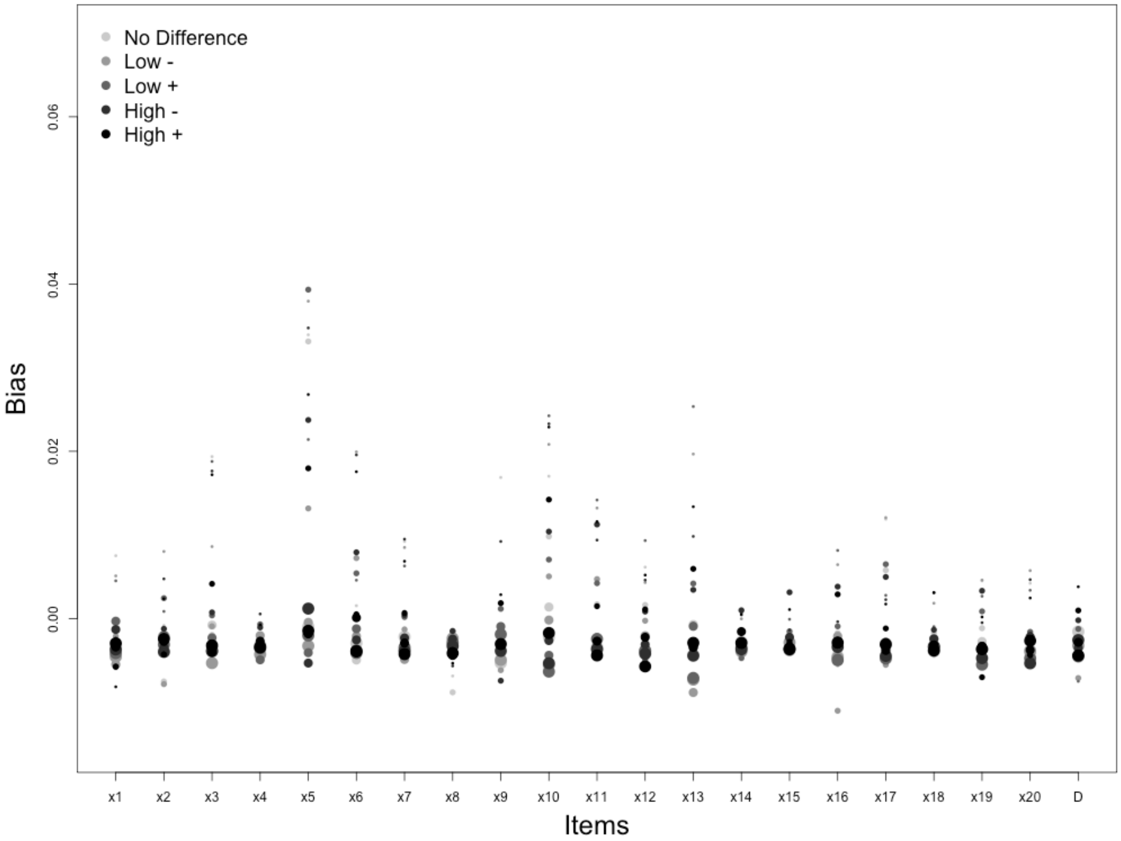

Figure A21: Plot of bias difficulty for the 20-item test under the ability-based scenario.

Note: The sample size is represented by the increase size in dots on the plot with the smallest dot representing N = 500 and the largest representing N = 10,000.

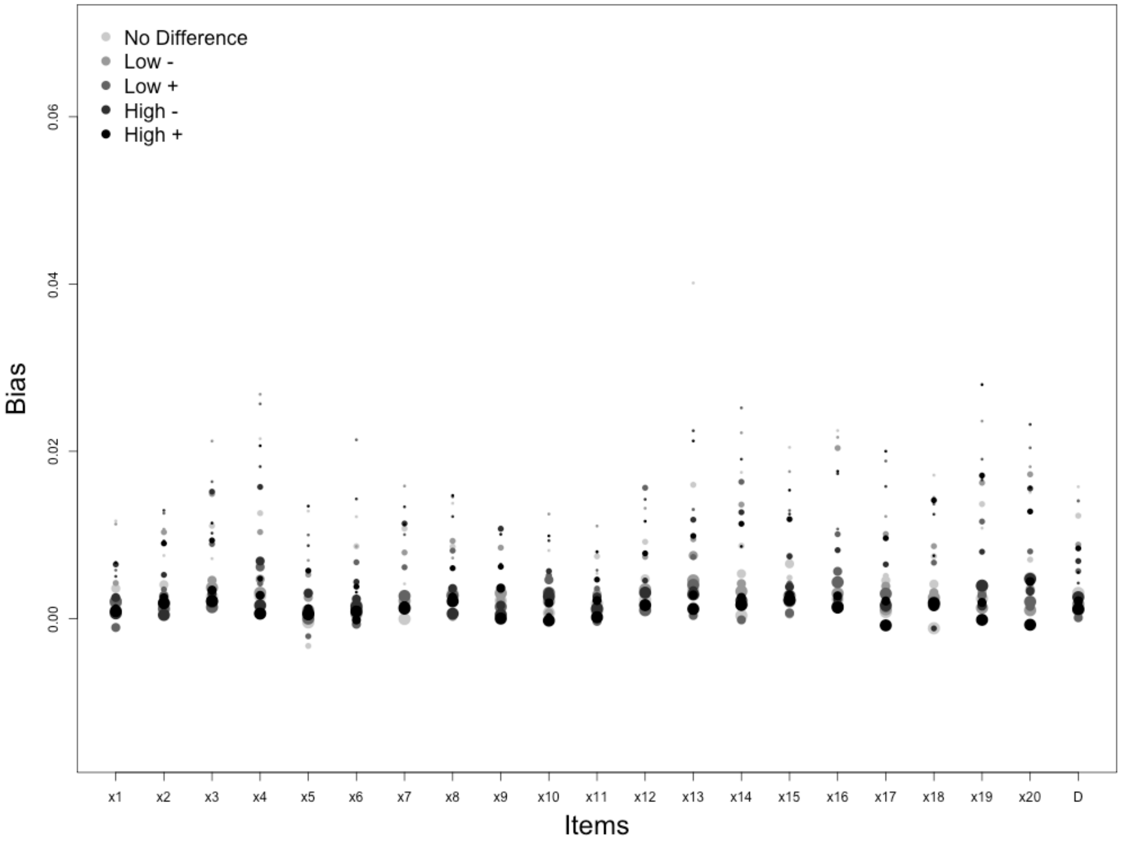

Figure A22: Plot of bias discrimination for the 20-item test under the ability-based scenario.

Note: The sample size is represented by the increase size in dots on the plot with the smallest dot representing N = 500 and the largest representing N = 10,000.

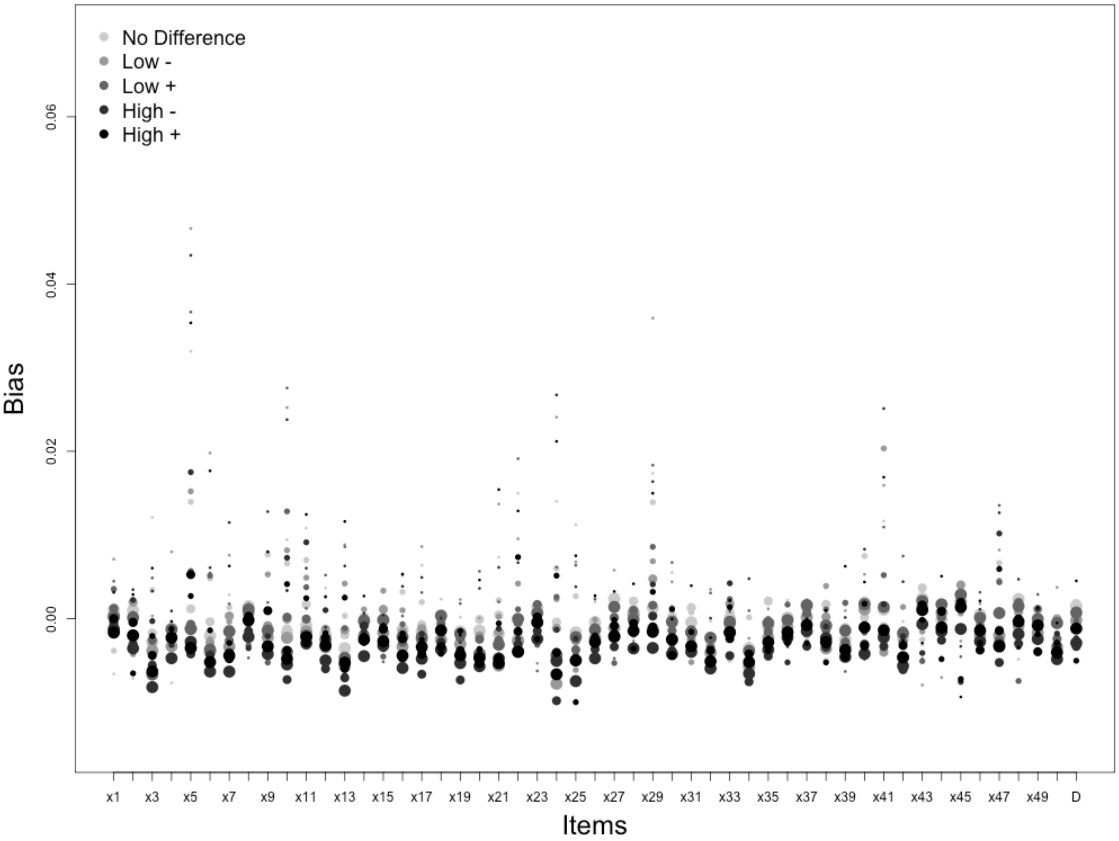

Figure A23: Plot of bias difficulty for the 50-item test under the ability-based scenario.

Note: The sample size is represented by the increase size in dots on the plot with the smallest dot representing N = 500 and the largest representing N = 10,000.

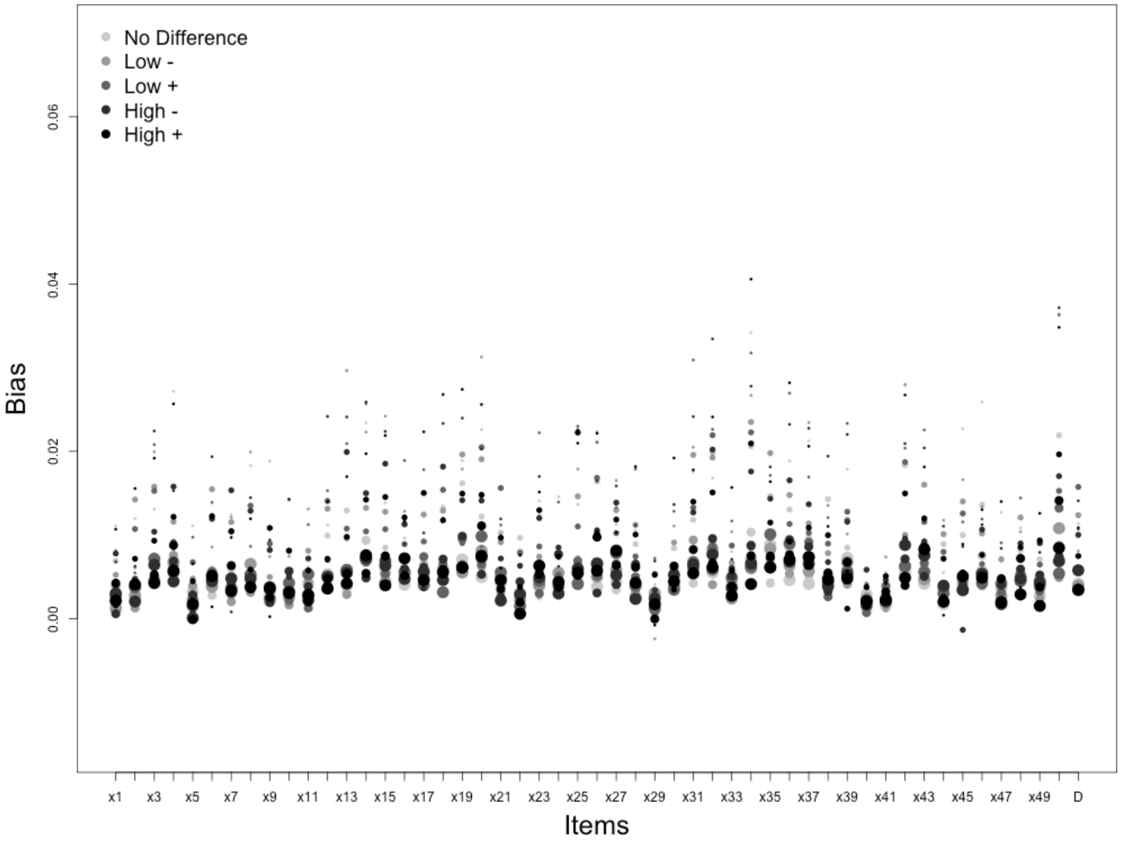

Figure A24: Plot of bias discrimination for the 50-item test under the ability-based scenario.

Note: The sample size is represented by the increase size in dots on the plot with the smallest dot representing N = 500 and the largest representing N = 10,000.

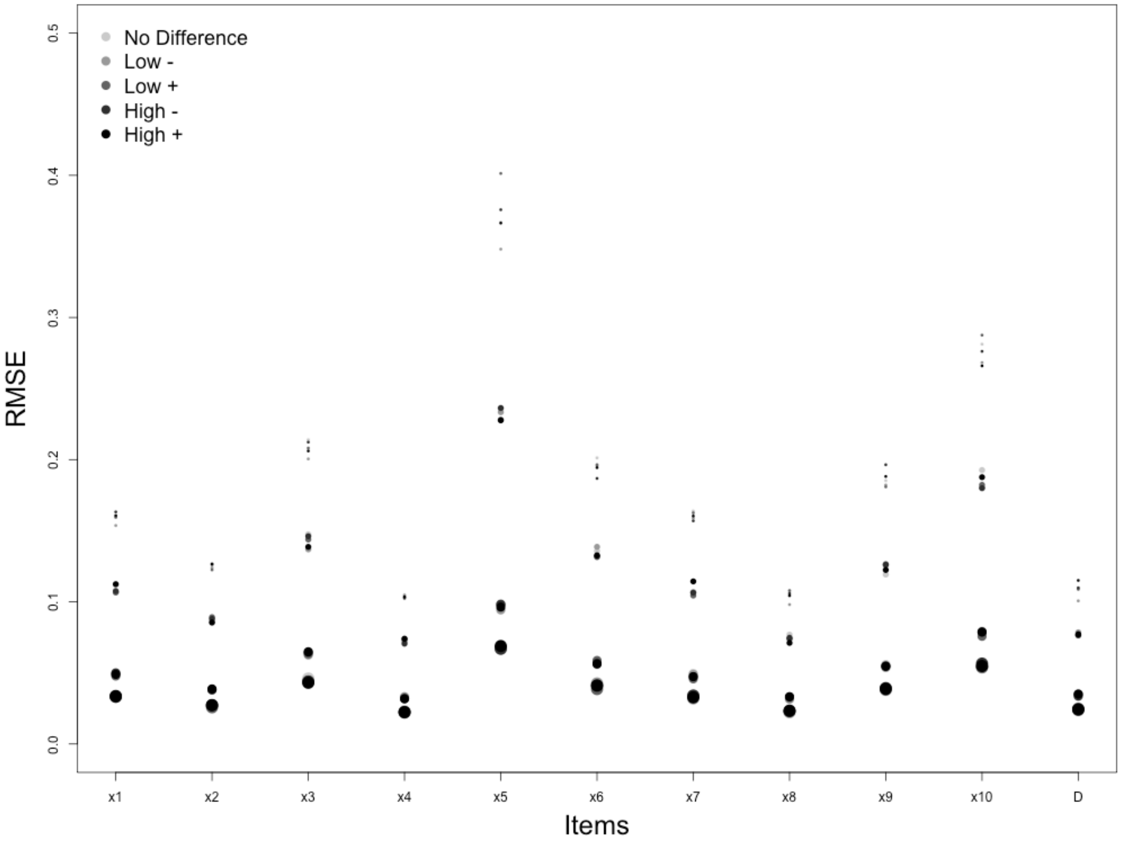

Figure A25: Plot of RMSE difficulty for the 10-item test under the ability-based scenario.

Note: The sample size is represented by the increase size in dots on the plot with the smallest dot representing N = 500 and the largest representing N = 10,000.

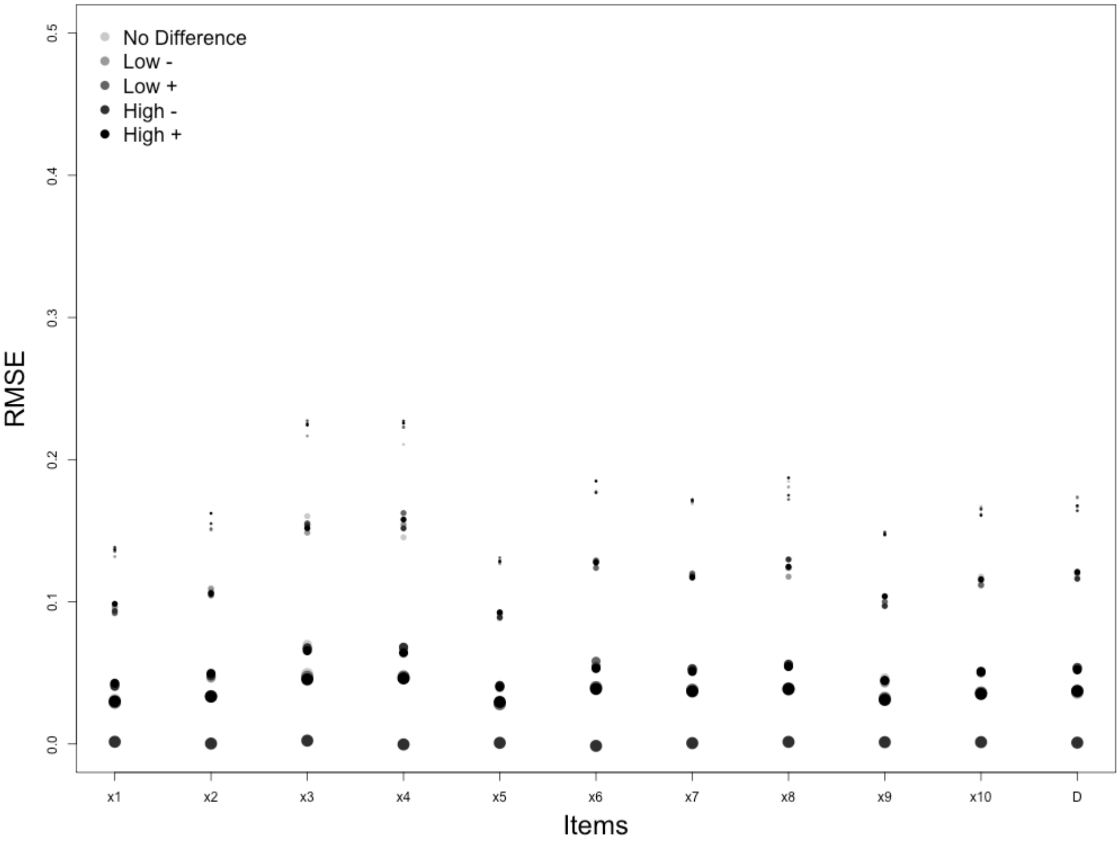

Figure A26: Plot of RMSE discrimination for the 10-item test under the ability-based scenario.

Note: The sample size is represented by the increase size in dots on the plot with the smallest dot representing N = 500 and the largest representing N = 10,000.

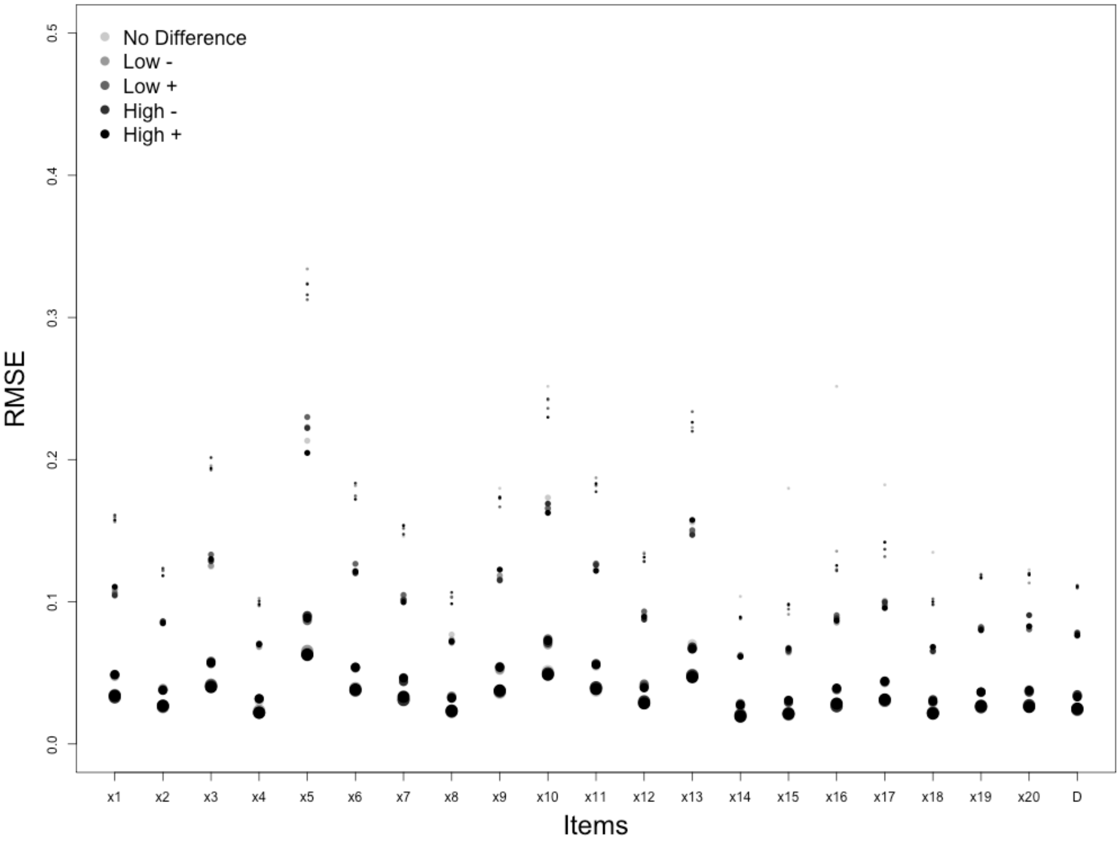

Figure A27: Plot of RMSE difficulty for the 20-item test under the ability-based scenario.

Note: The sample size is represented by the increase size in dots on the plot with the smallest dot representing N = 500 and the largest representing N = 10,000.

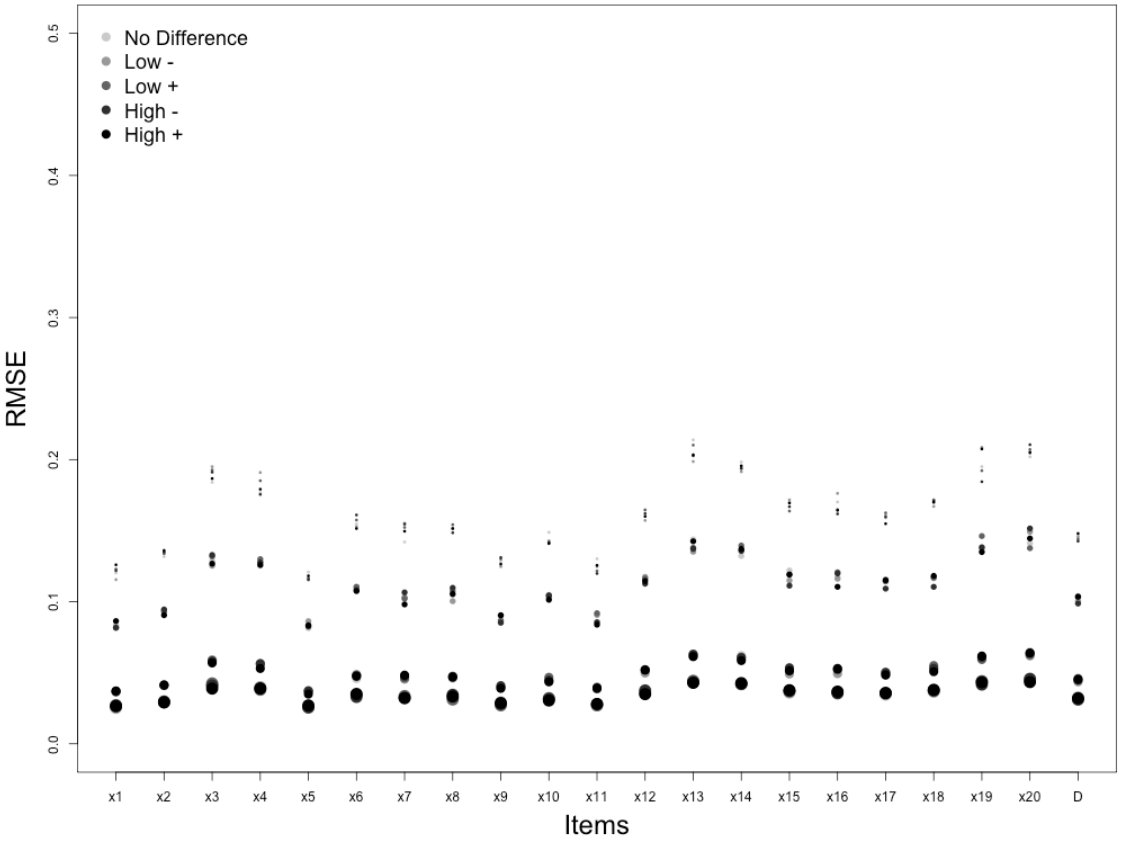

Figure A28: Plot of RMSE discrimination for the 20-item test under the ability-based scenario.

Note: The sample size is represented by the increase size in dots on the plot with the smallest dot representing N = 500 and the largest representing N = 10,000.

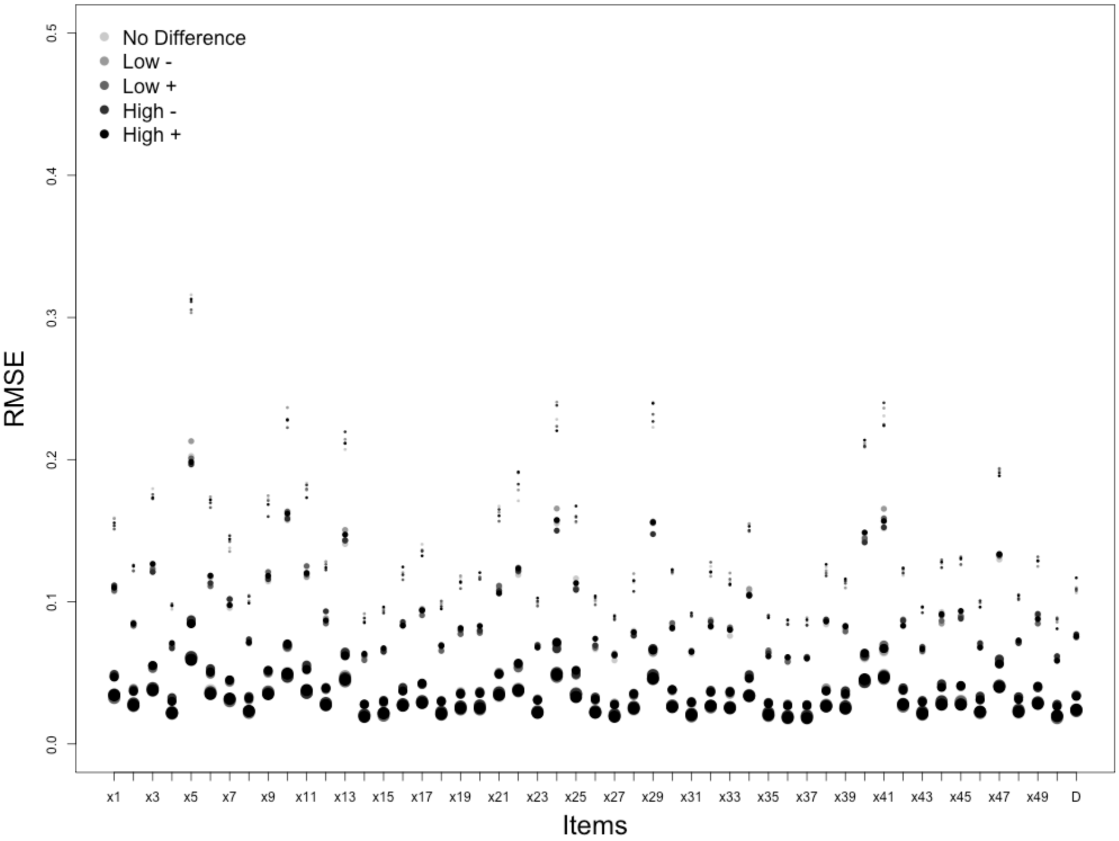

Figure A29: Plot of RMSE difficulty for the 50-item test under the ability-based scenario.

Note: The sample size is represented by the increase size in dots on the plot with the smallest dot representing N = 500 and the largest representing N = 10,000.

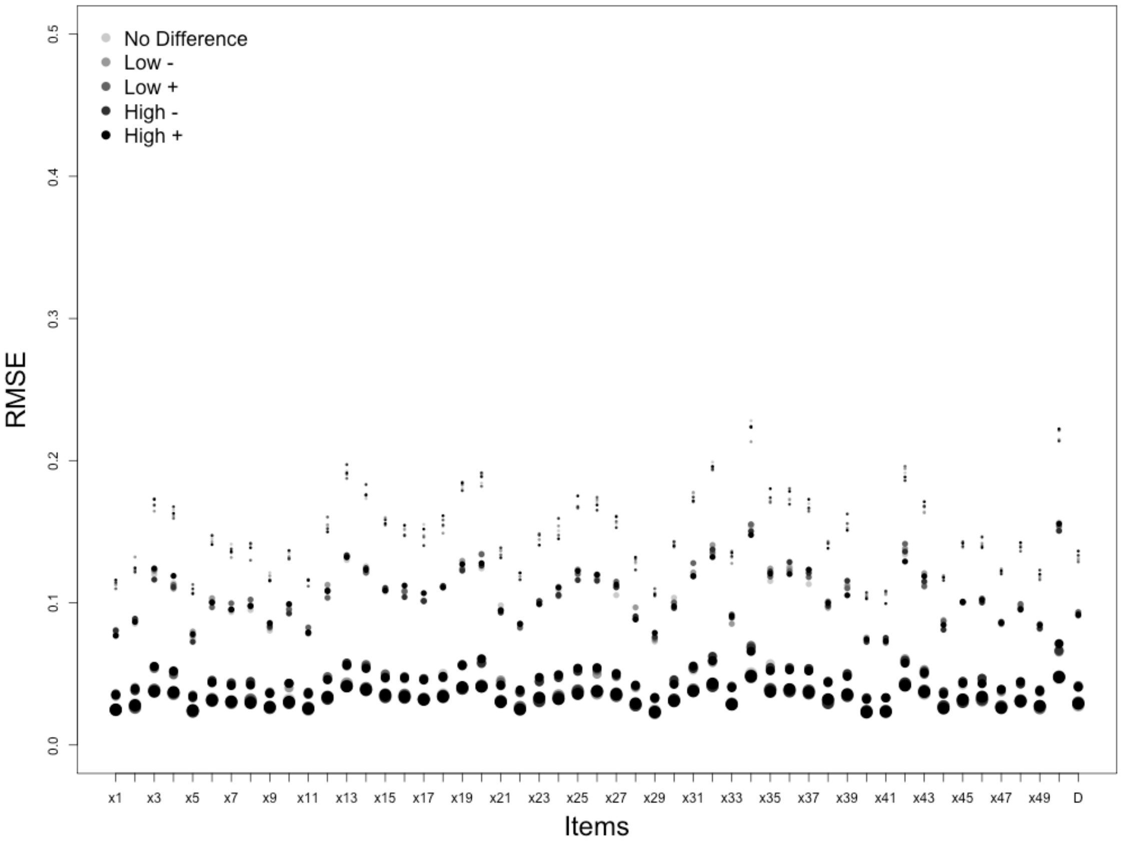

Figure A30: Plot of RMSE discrimination for the 50-item test under the ability-based scenario.

Note: The sample size is represented by the increase size in dots on the plot with the smallest dot representing N = 500 and the largest representing N = 10,000.

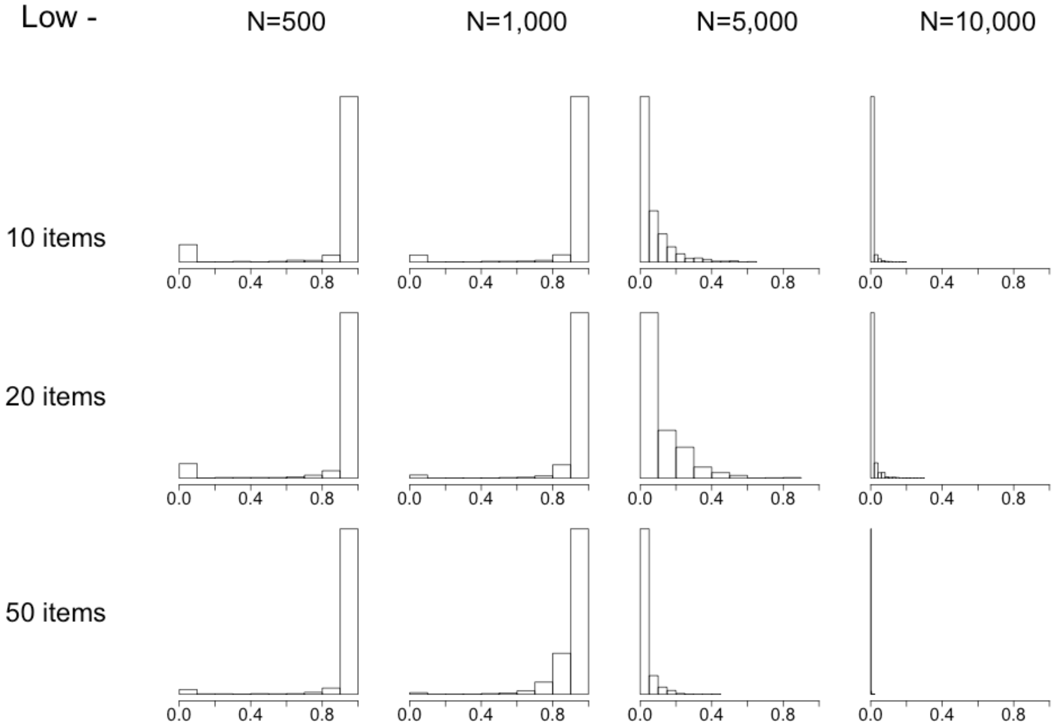

Figure A31: Histograms of the *p*-values from the IRT method under the cut-score scenario for the Low - outcome condition.

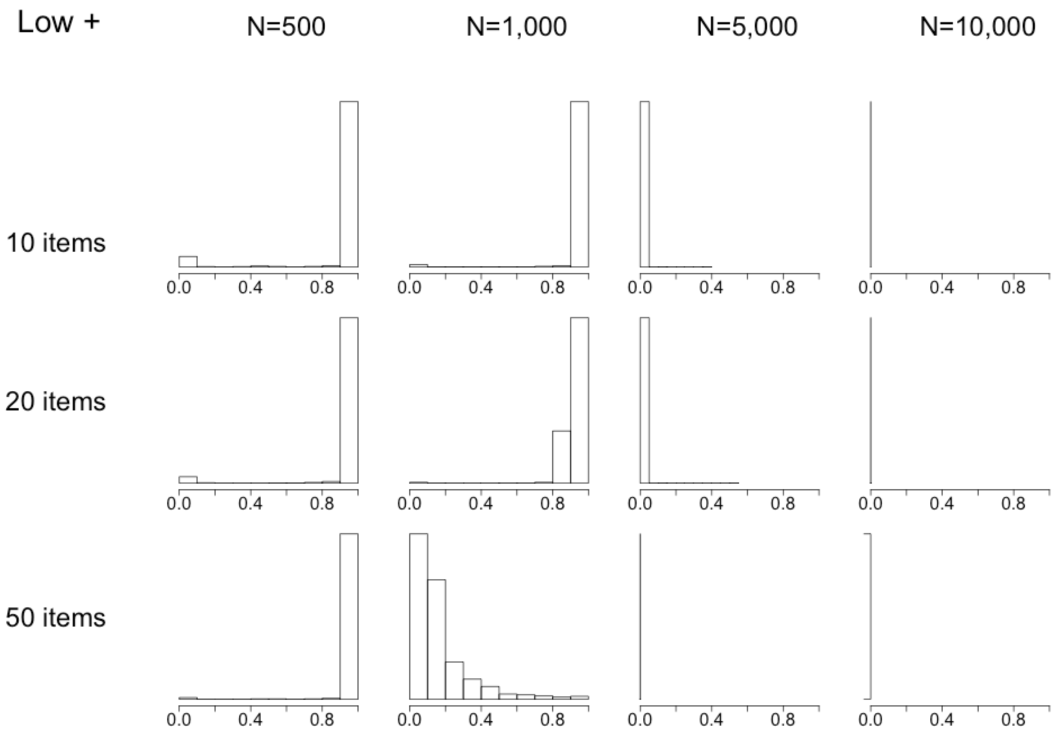

Figure A32: Histograms of the *p*-values from the IRT method under the cut-score scenario for the Low + outcome condition.

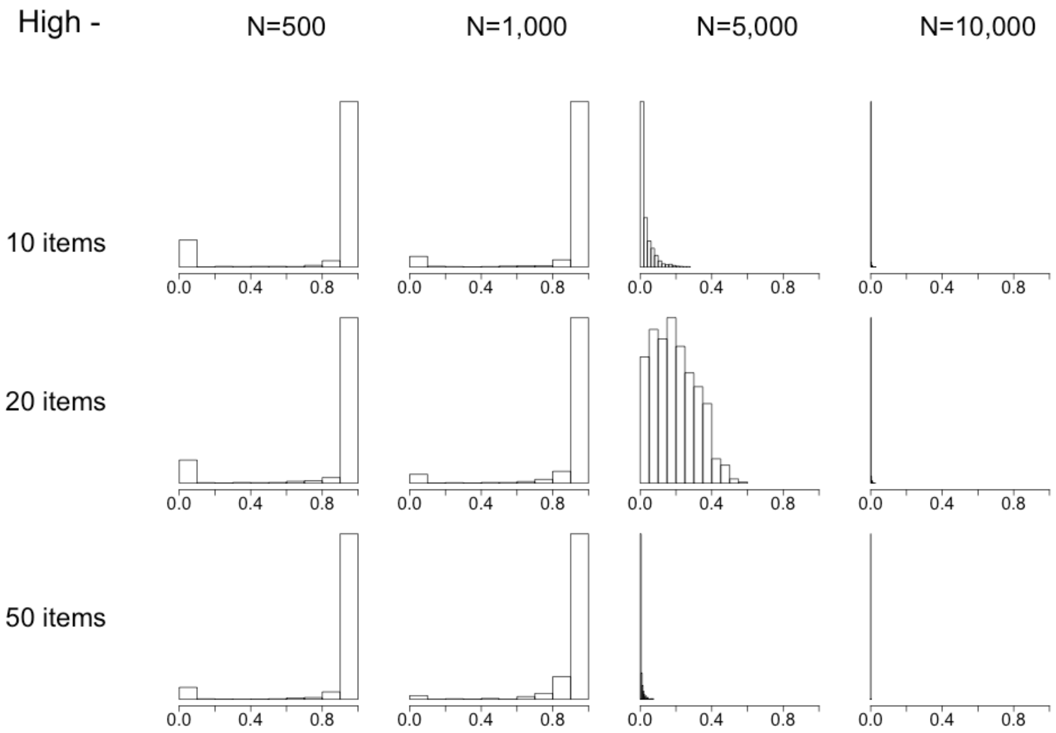

Figure A33: Histograms of the *p*-values from the IRT method under the cut-score scenario for the High - outcome condition.

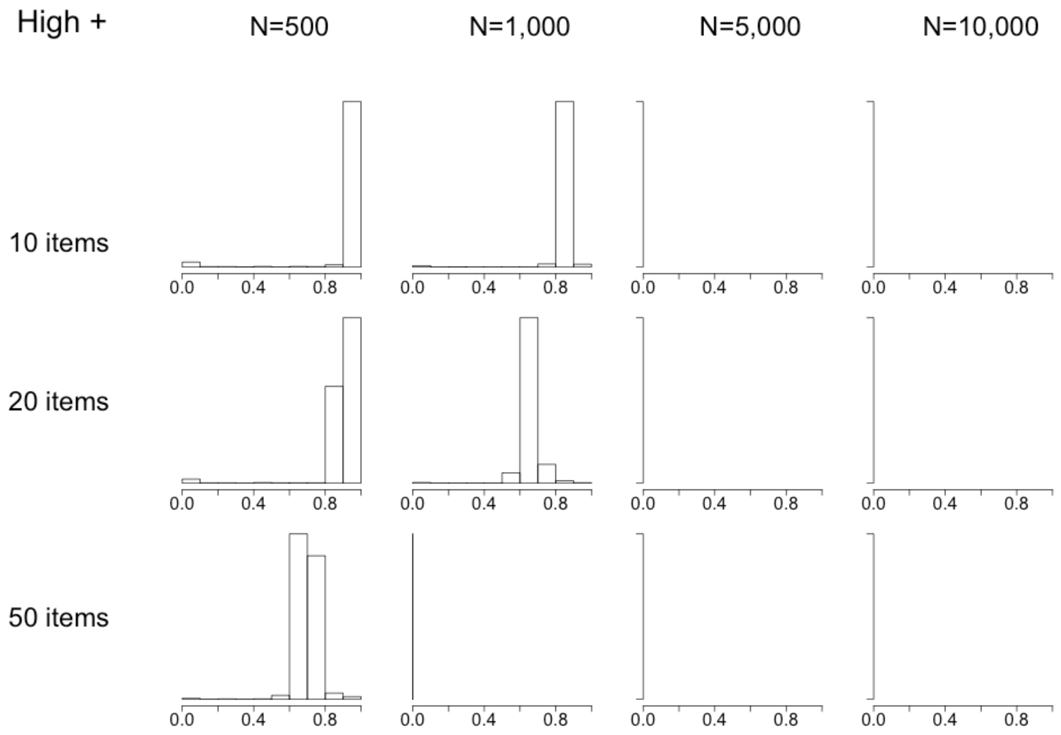

Figure A34: Histograms of the *p*-values from the IRT method under the cut-score scenario for the High + outcome condition.

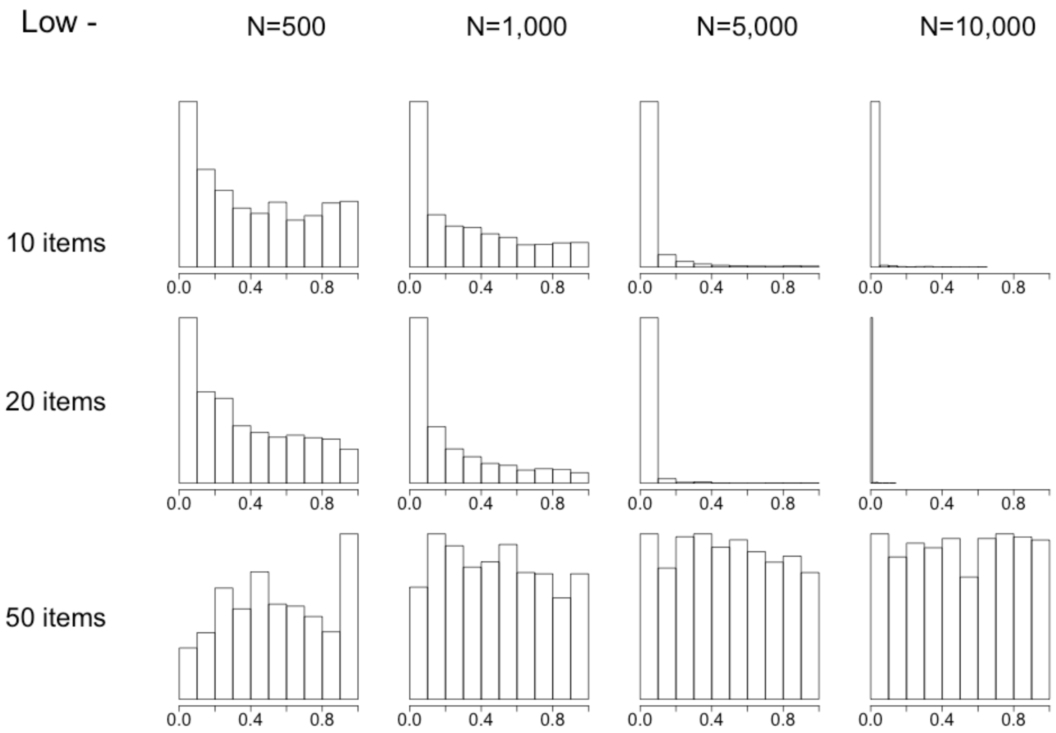

Figure A35: Histograms of the *p*-values from the LR method under the cut-score scenario for the Low - outcome condition.

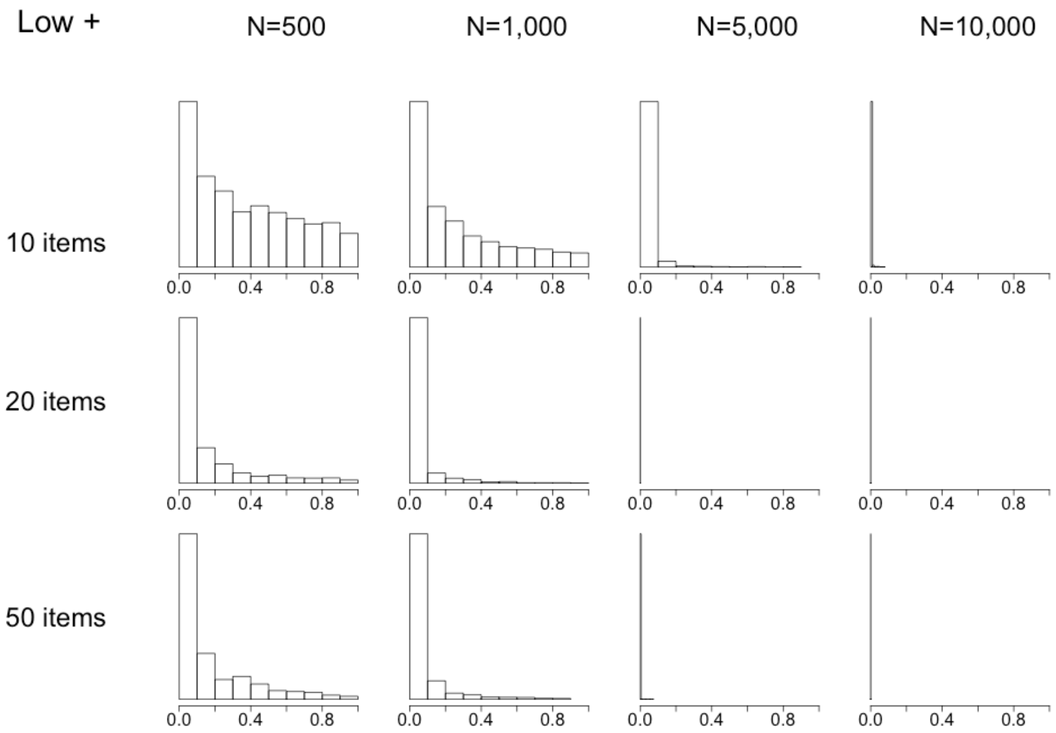

Figure A36: Histograms of the *p*-values from the LR method under the cut-score scenario for the Low + outcome condition.

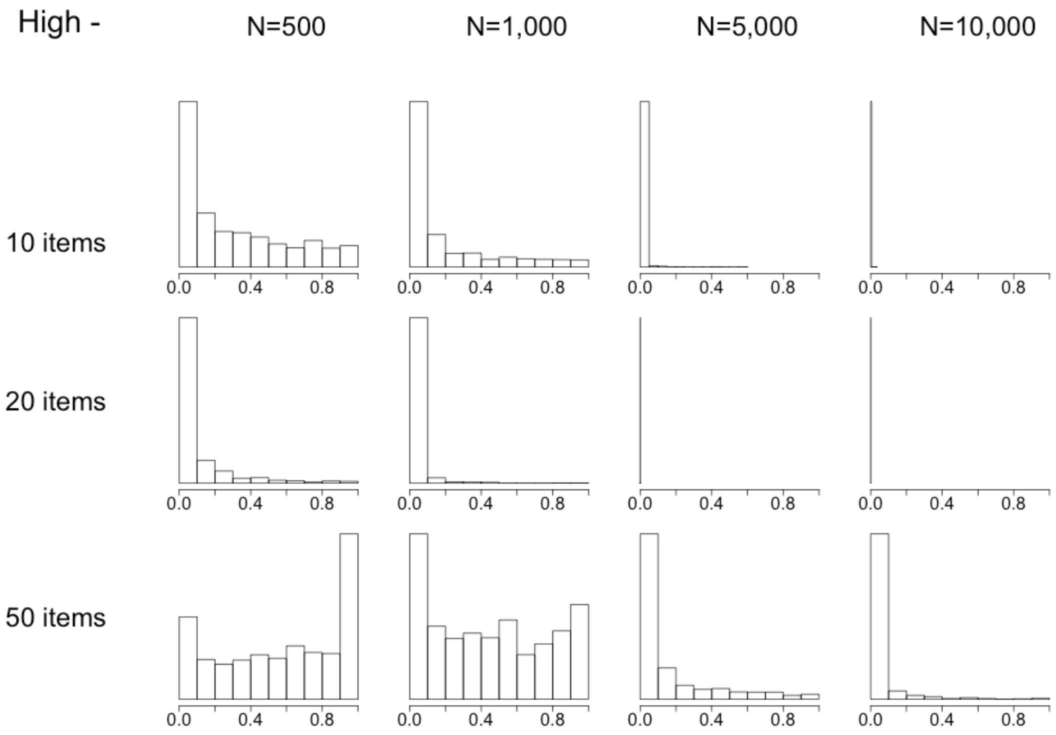

Figure A37: Histograms of the *p*-values from the LR method under the cut-score scenario for the High - outcome condition.

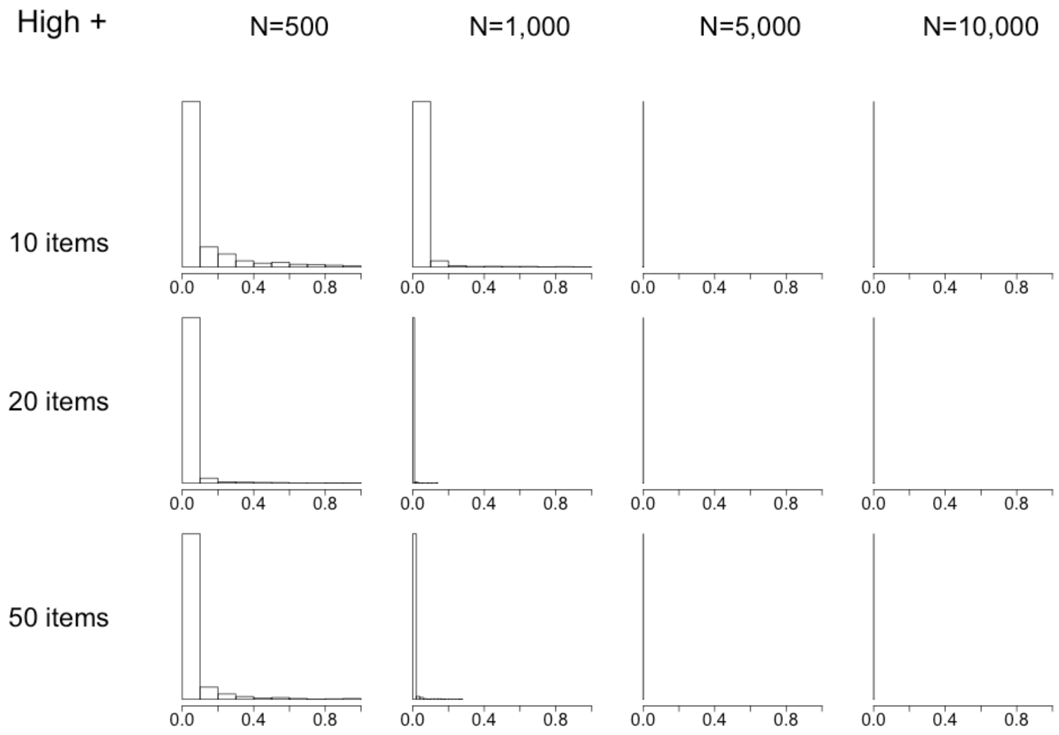

Figure A38: Histograms of the *p*-values from the LR method under the cut-score scenario for the High + outcome condition.

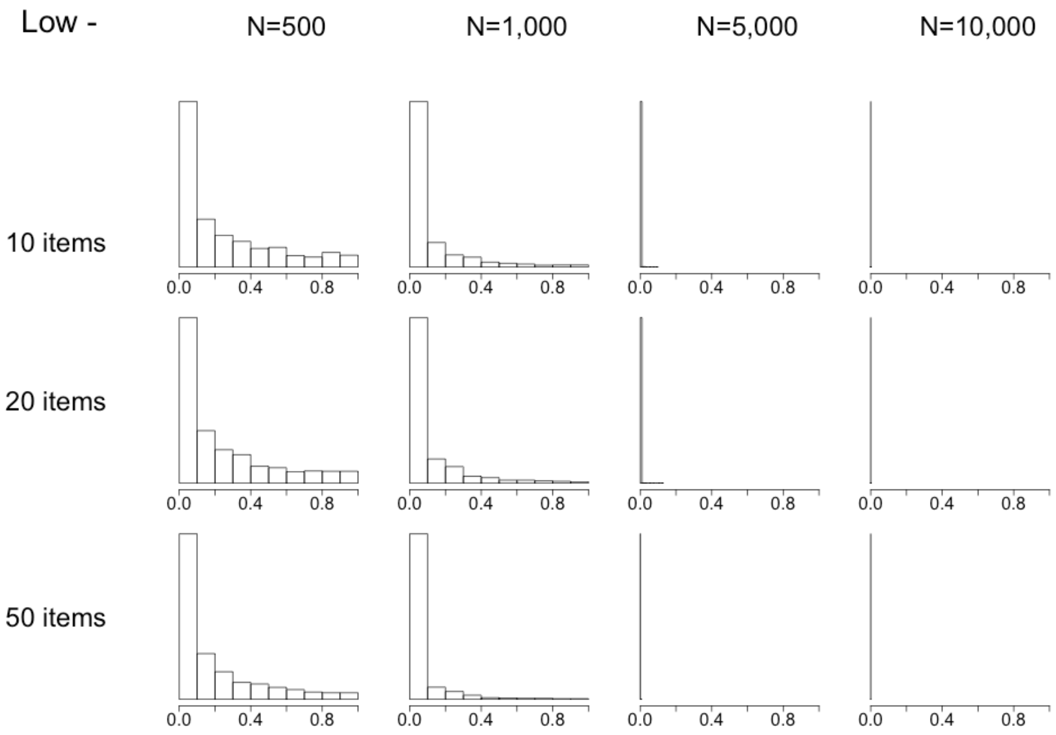

Figure A39: Histograms of the *p*-values from the RD method under the cut-score scenario for the Low - outcome condition.

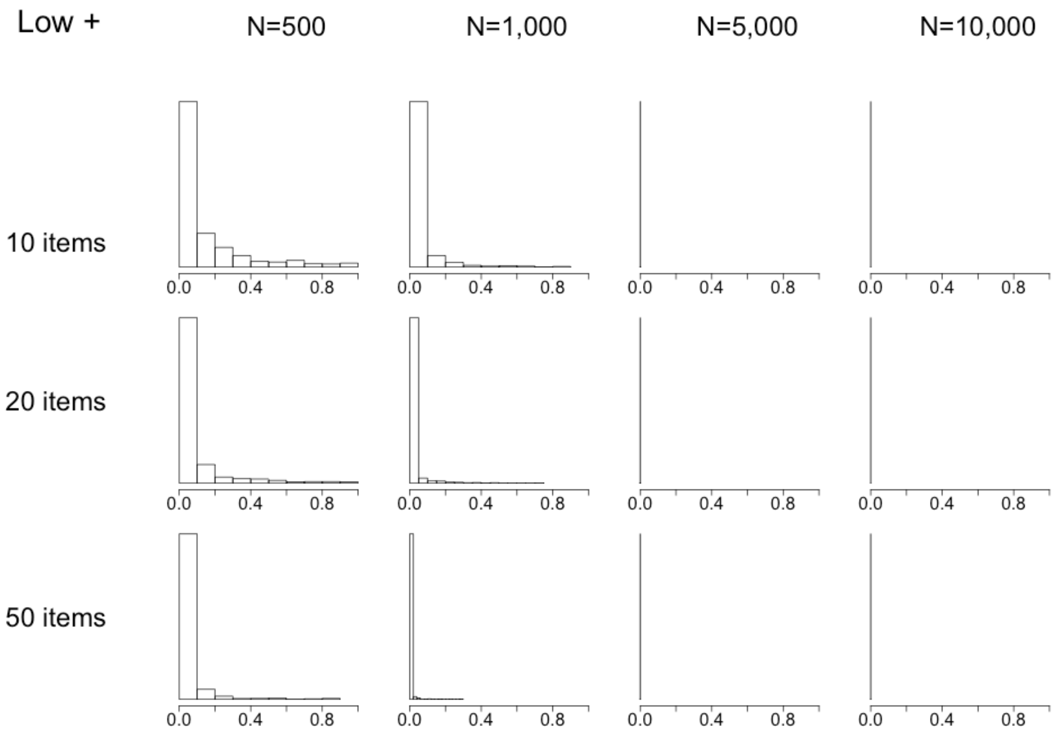

Figure A40: Histograms of the *p*-values from the RD method under the cut-score scenario for the Low + outcome condition.

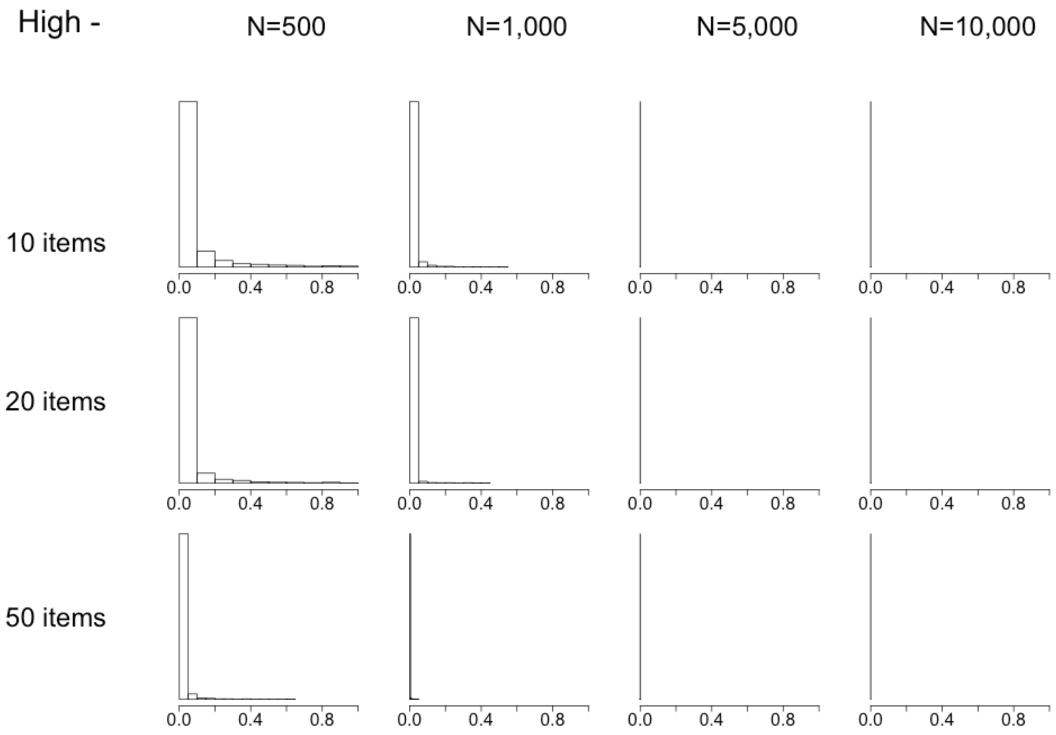

Figure A41: Histograms of the *p*-values from the RD method under the cut-score scenario for the High - outcome condition.

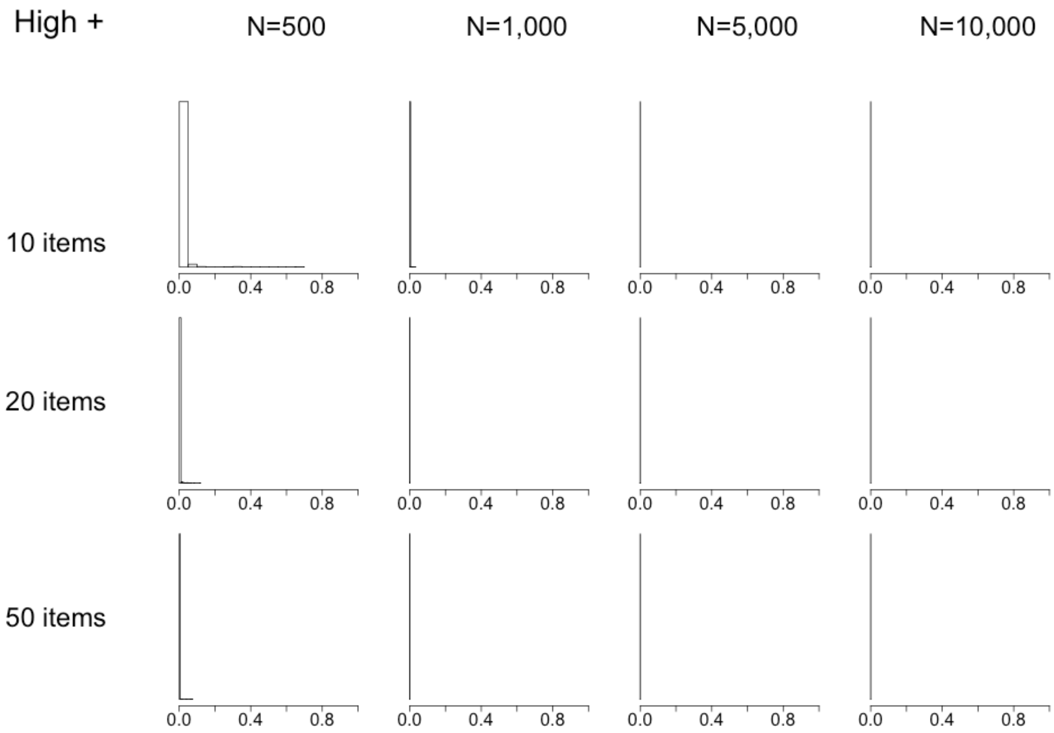

Figure A42: Histograms of the *p*-values from the RD method under the cut-score scenario for the High + outcome condition.

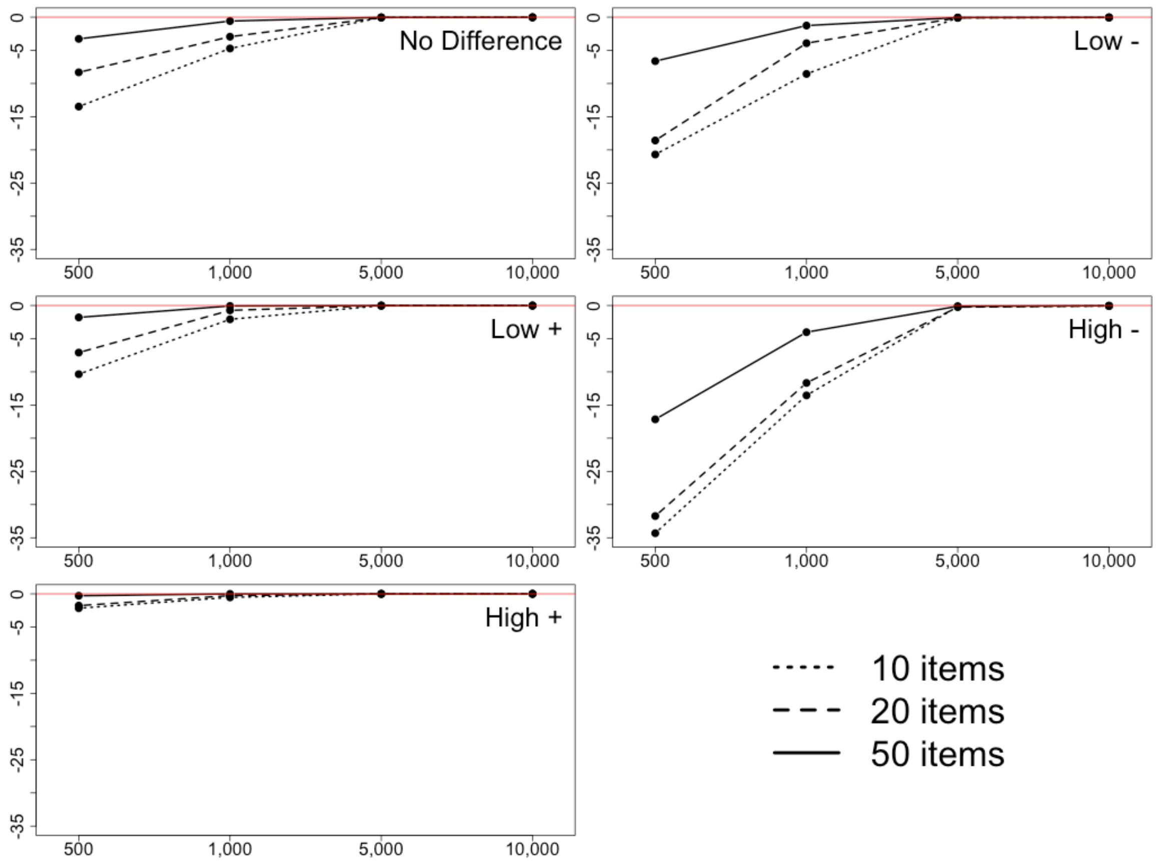

Figure A43: Profile plot for the bias of the difficulty parameter of the outcome variable for the control group under the cut-score scenario.

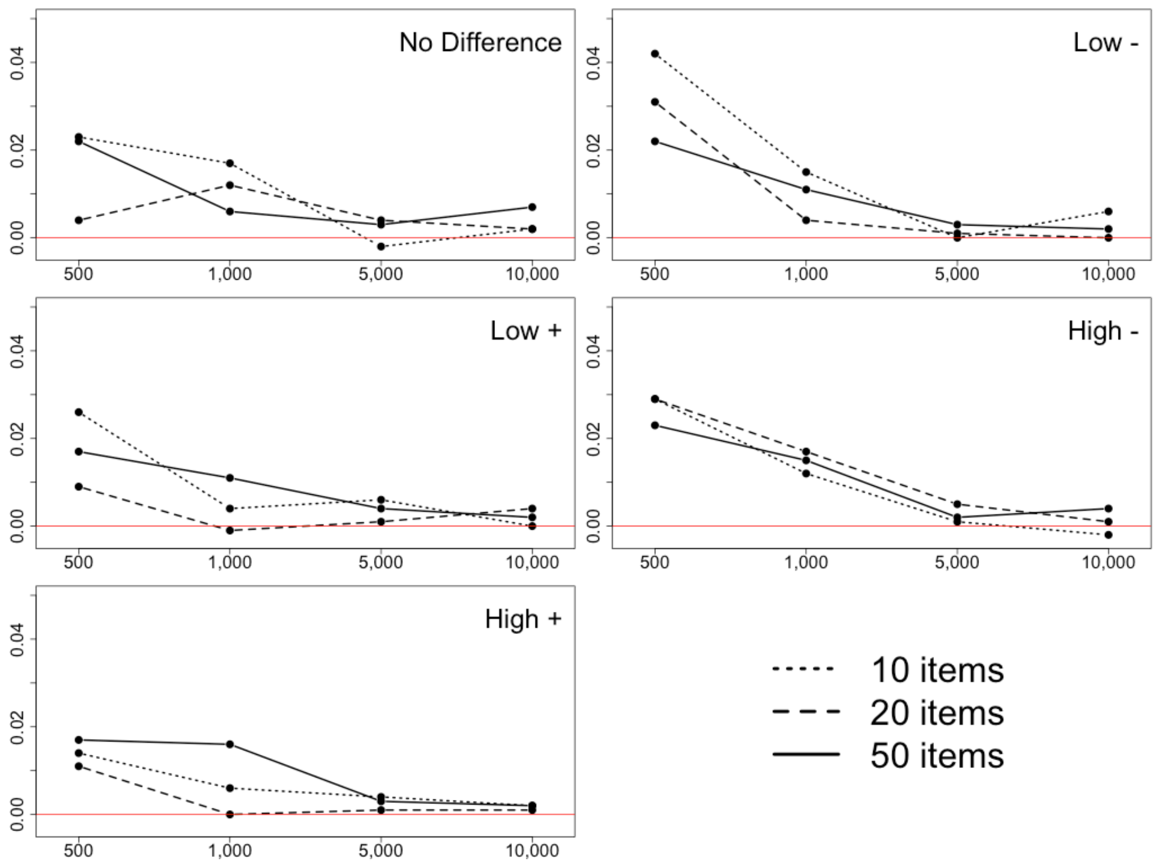

Figure A44: Profile plot for the bias of the discrimination parameter of the outcome variable for the treatment group under the cut-score scenario.

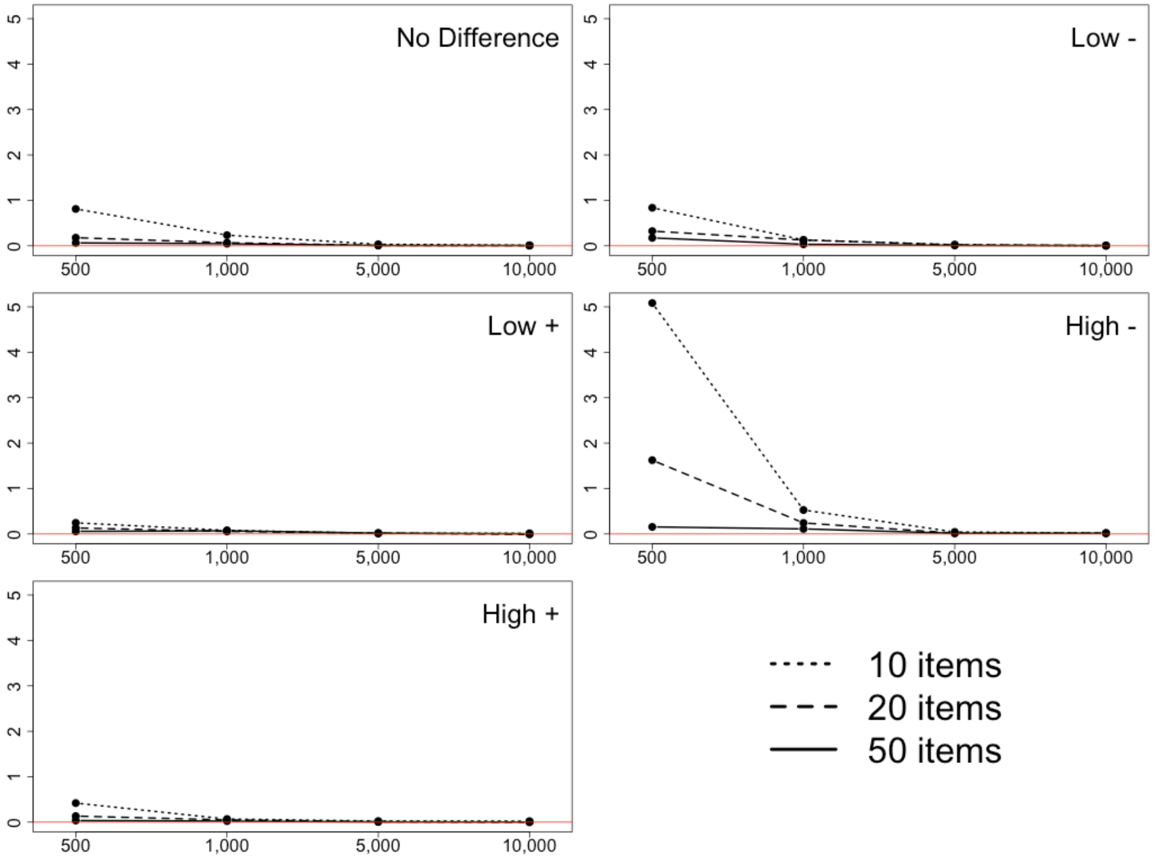

Figure A45: Profile plot for the bias of the discrimination parameter of the outcome variable for the control group under the cut-score scenario.

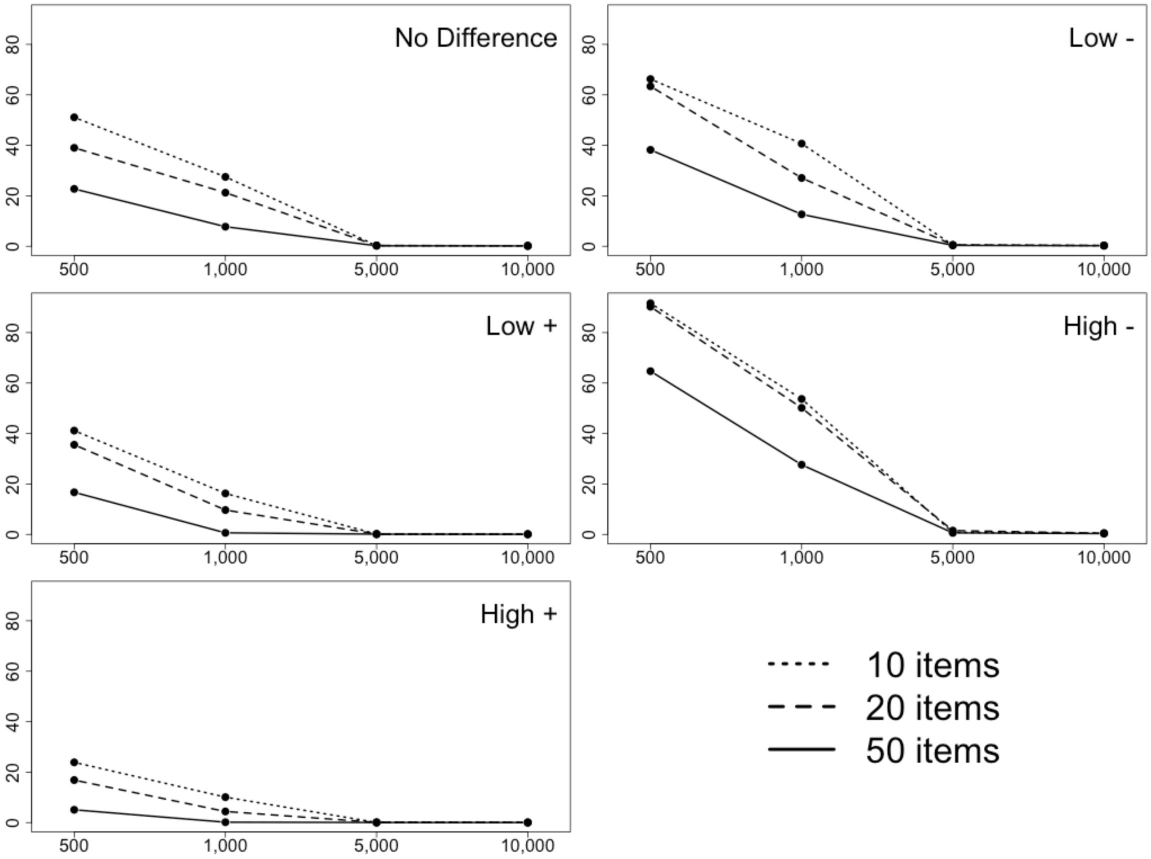

Figure A46: Profile plot for the RMSE of the difficulty parameter of the outcome variable for the control group under the cut-score scenario.

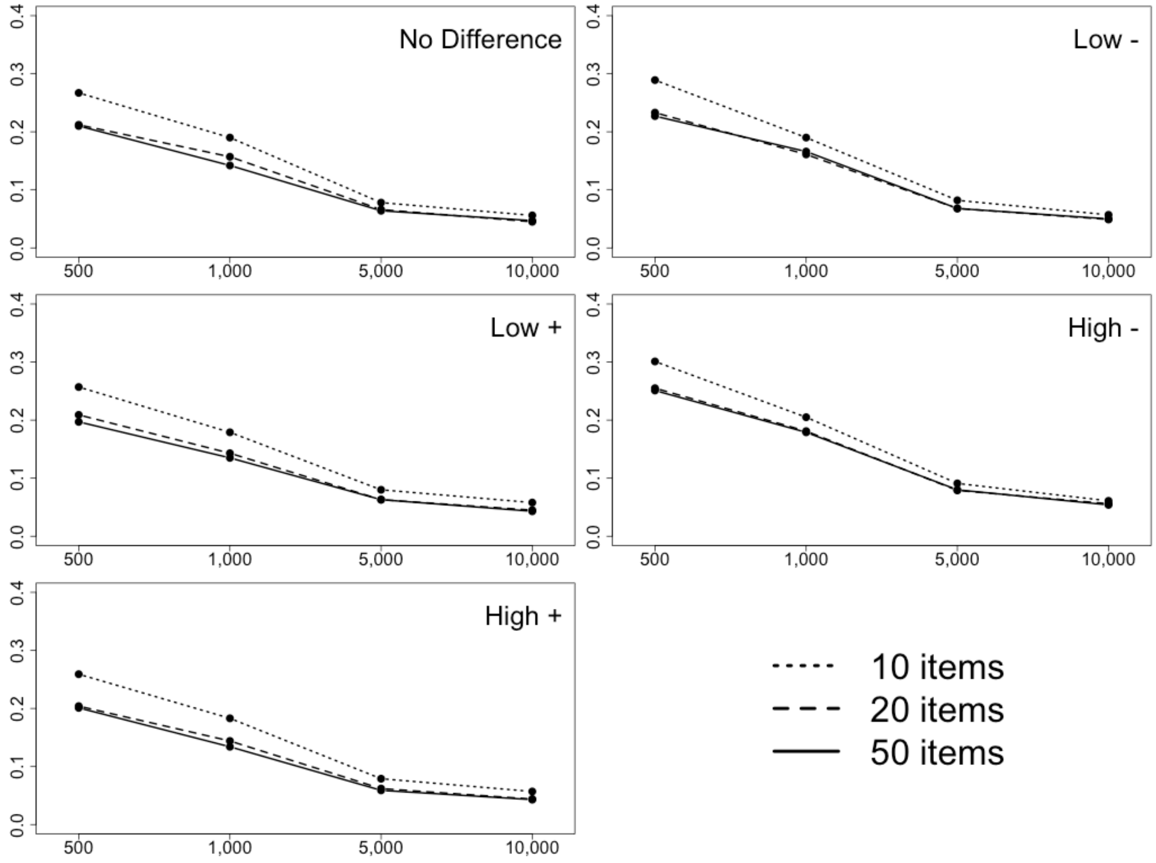

Figure A47: Profile plot for the RMSE of the discrimination parameter of the outcome variable for the treatment group under the cut-score scenario.

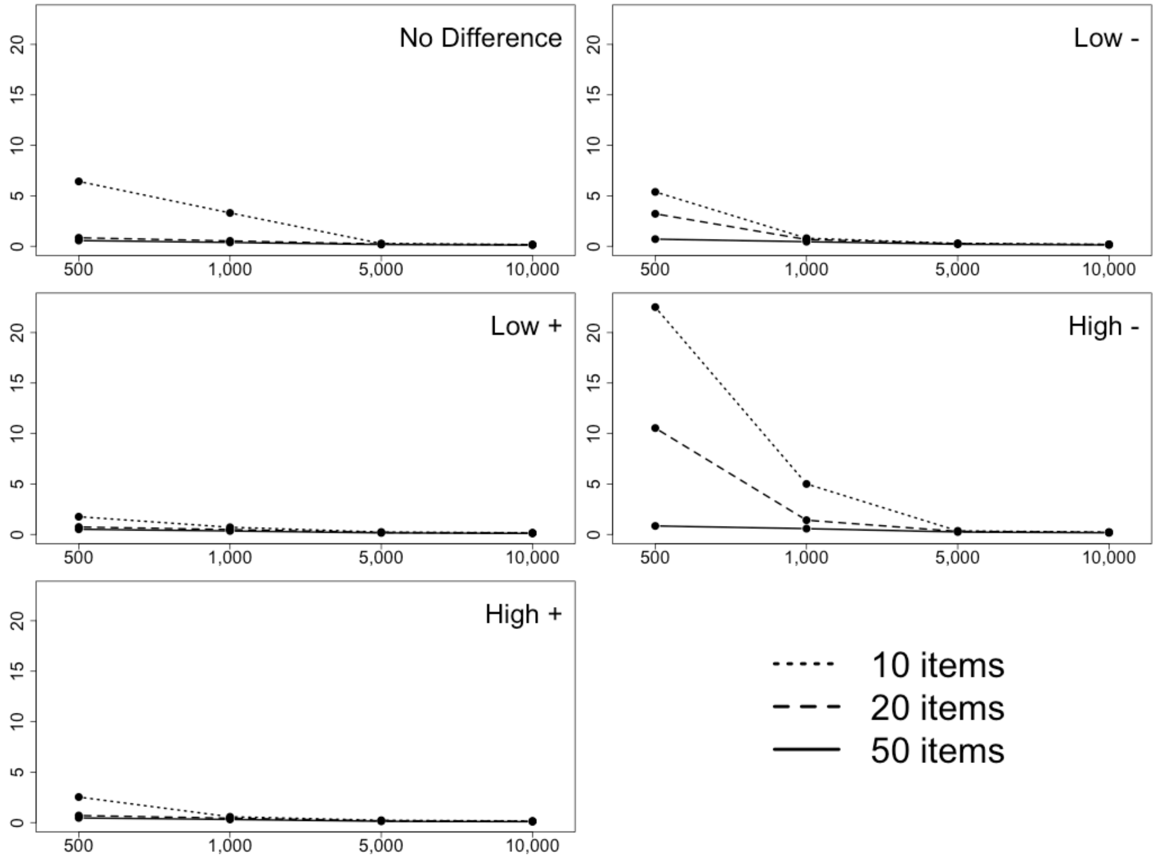

Figure A48: Profile plot for the RMSE of the discrimination parameter of the outcome variable for the control group under the cut-score scenario.

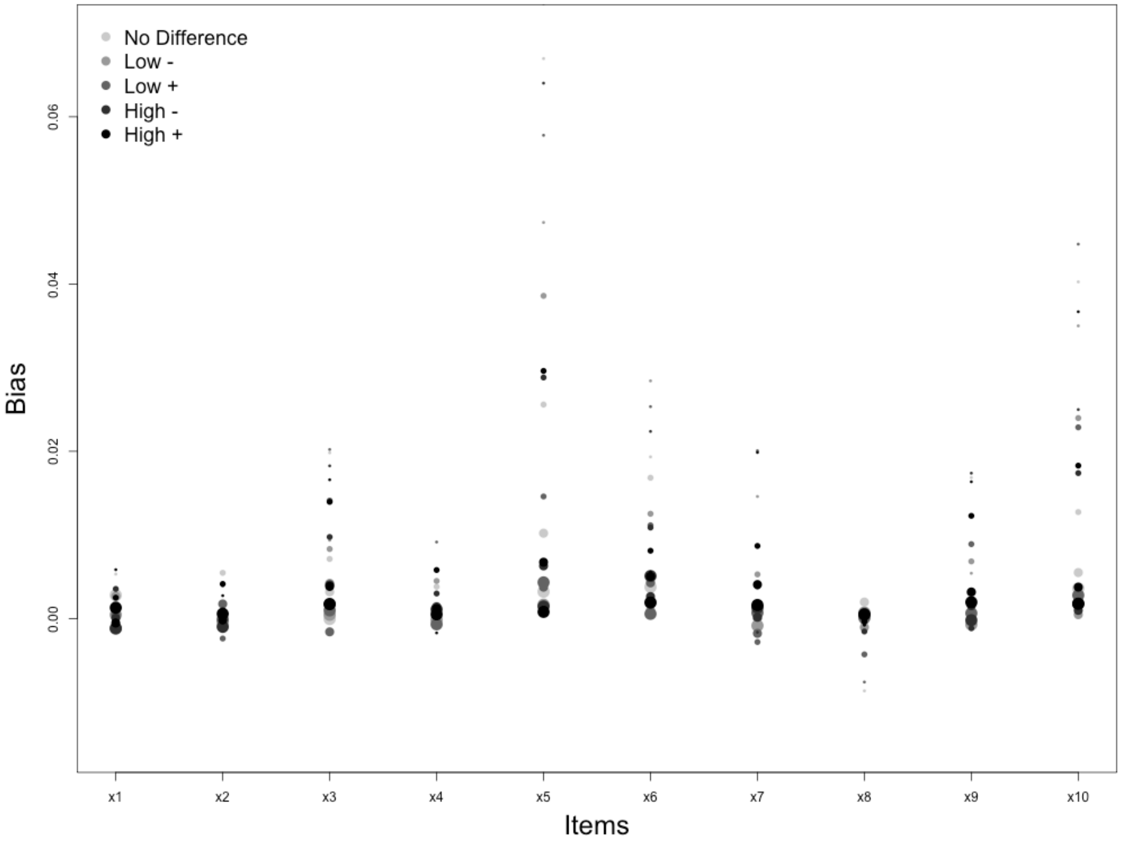

Figure A49: Plot of bias of the difficulty for the 10-item test under the cut-score scenario.

Note: The sample size is represented by the increase size in dots on the plot with the smallest dot representing N = 500 and the largest representing N = 10,000.

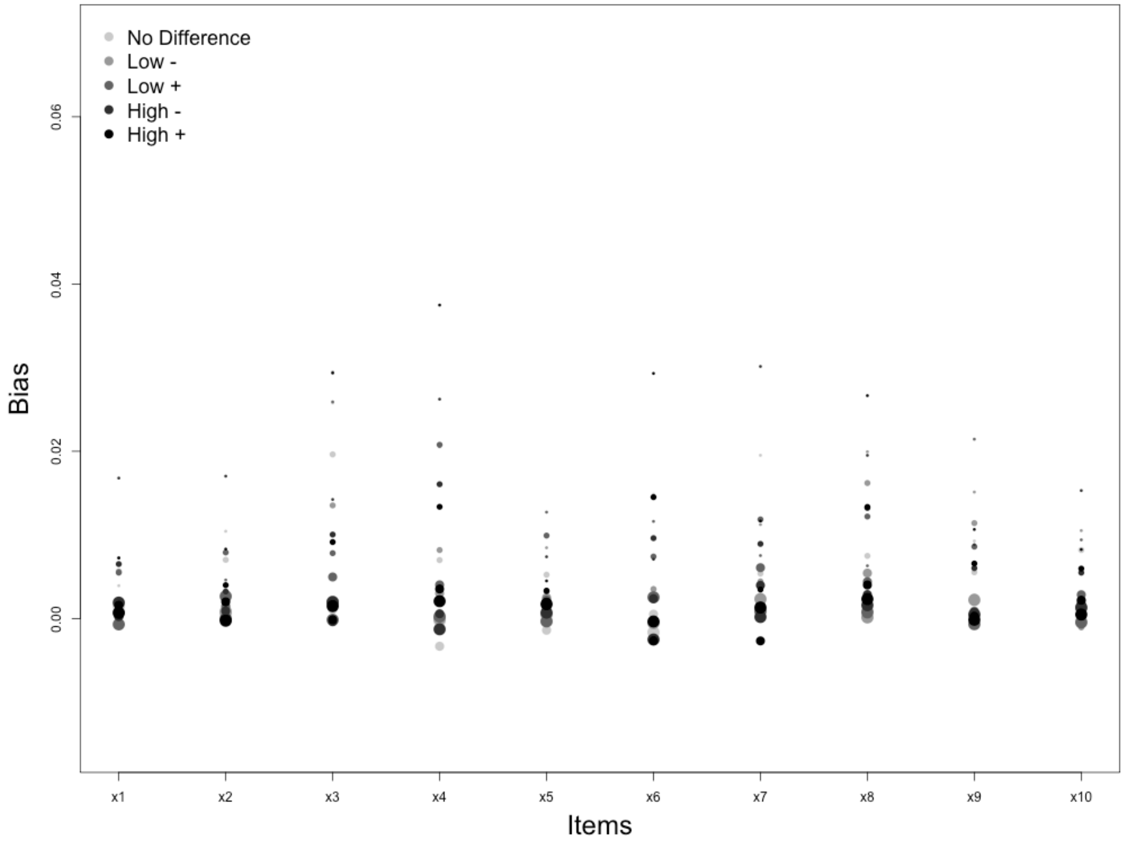

Figure A50: Plot of bias discrimination for the 10-item test under the cut-score scenario.

Note: The sample size is represented by the increase size in dots on the plot with the smallest dot representing N = 500 and the largest representing N = 10,000.

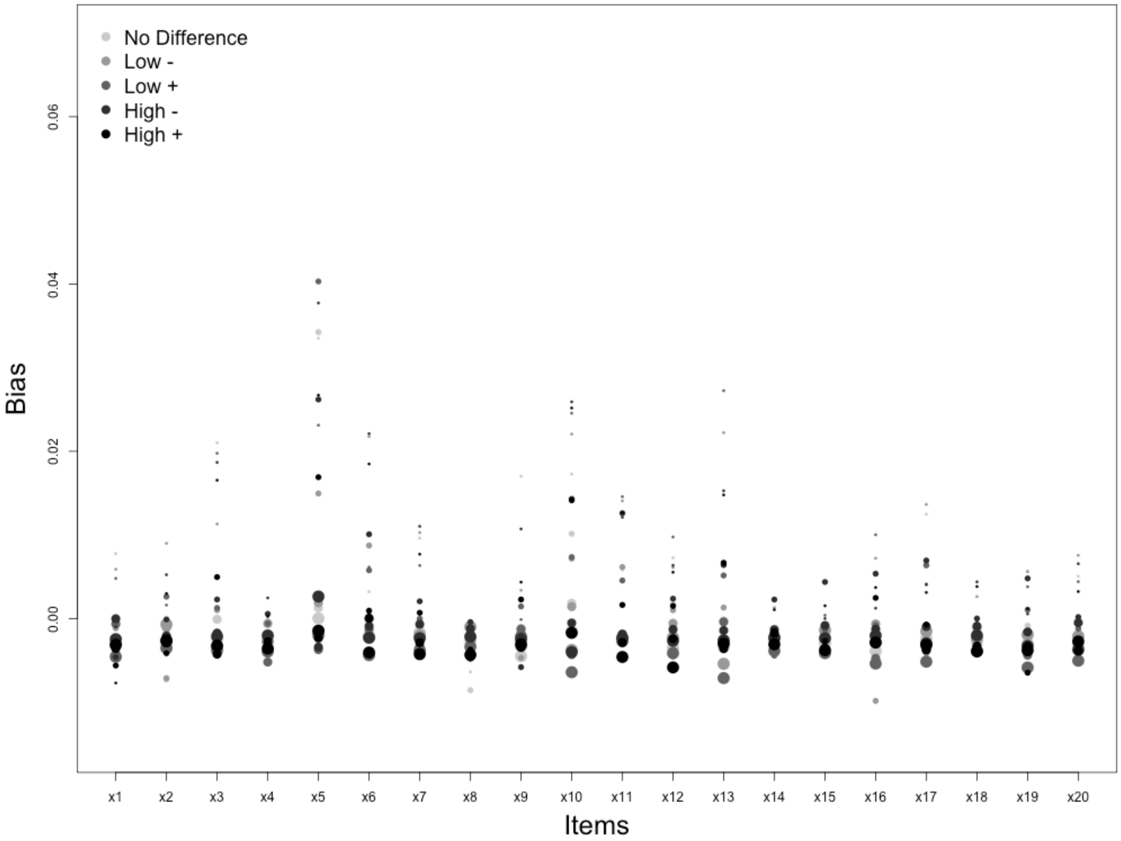

Figure A51: Plot of bias difficulty for the 20-item test under the cut-score scenario.

Note: The sample size is represented by the increase size in dots on the plot with the smallest dot representing N = 500 and the largest representing N = 10,000.

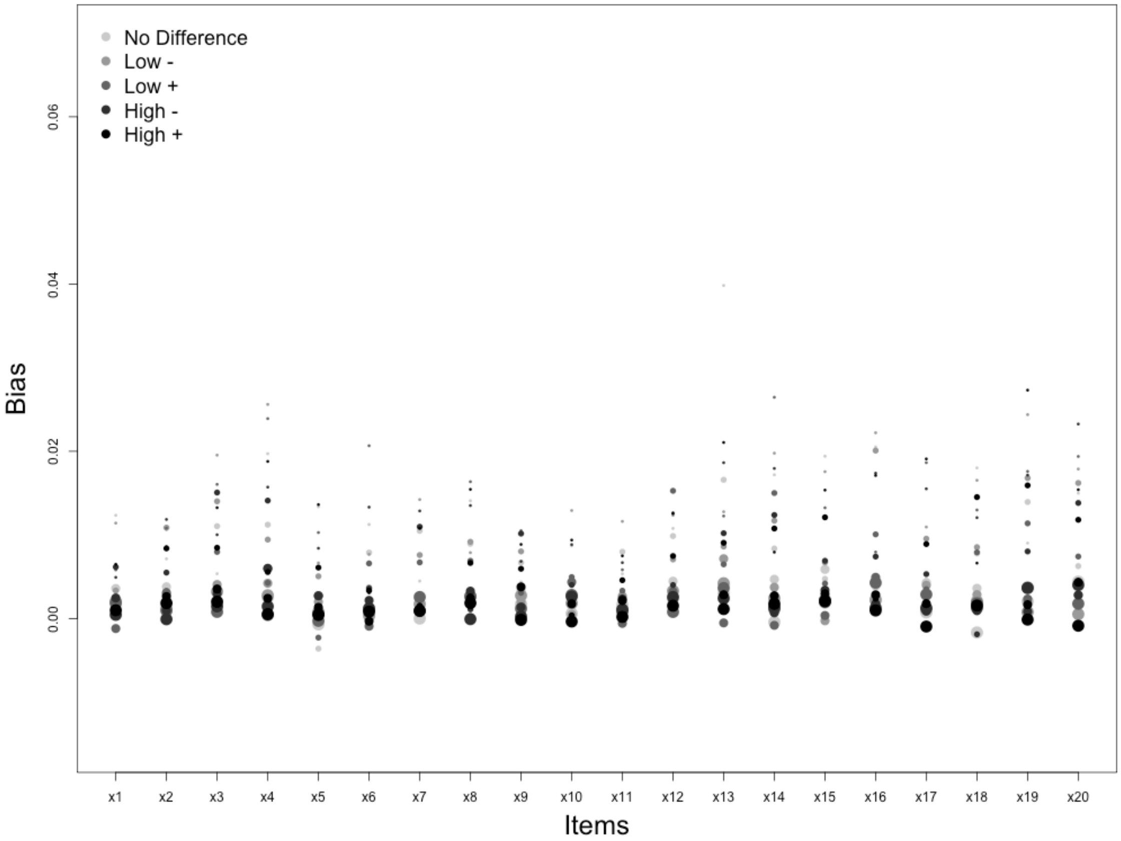

Figure A52: Plot of bias discrimination for the 20-item test under the cut-score scenario.

Note: The sample size is represented by the increase size in dots on the plot with the smallest dot representing N = 500 and the largest representing N = 10,000.

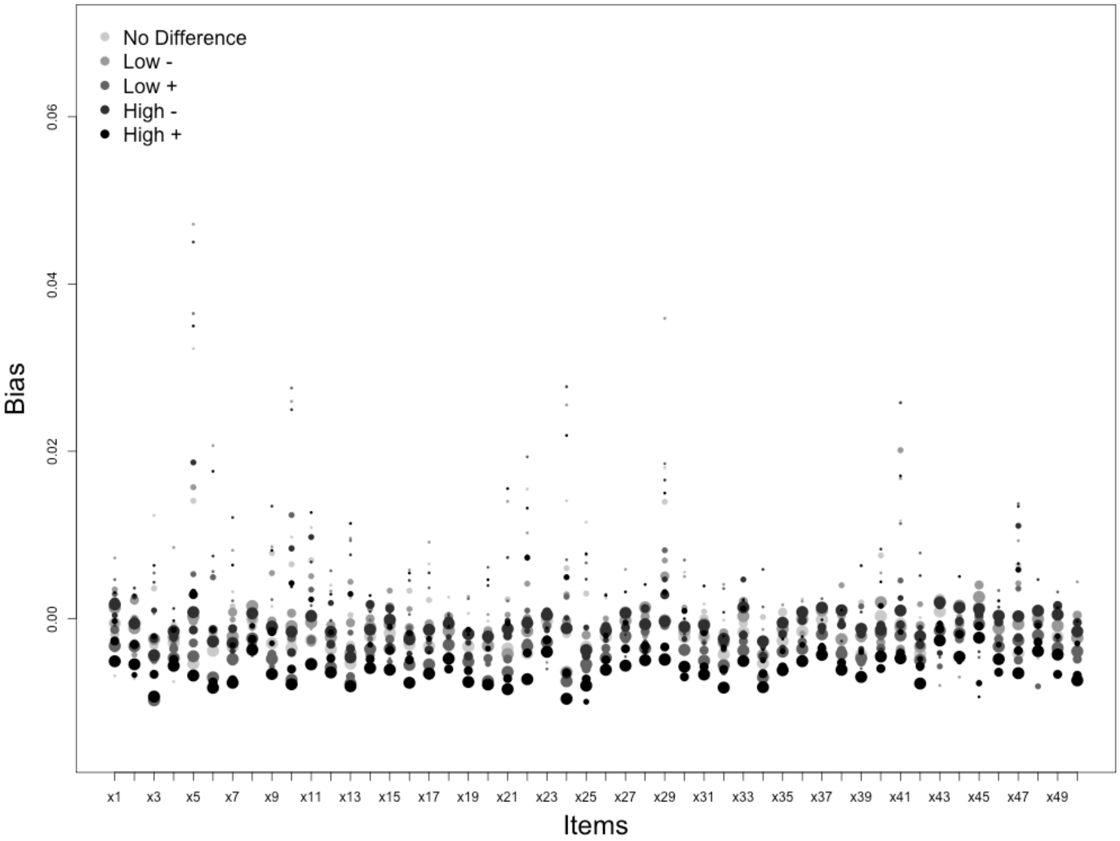

Figure A53: Plot of bias difficulty for the 50-item test under the cut-score scenario.

Note: The sample size is represented by the increase size in dots on the plot with the smallest dot representing N = 500 and the largest representing N = 10,000.

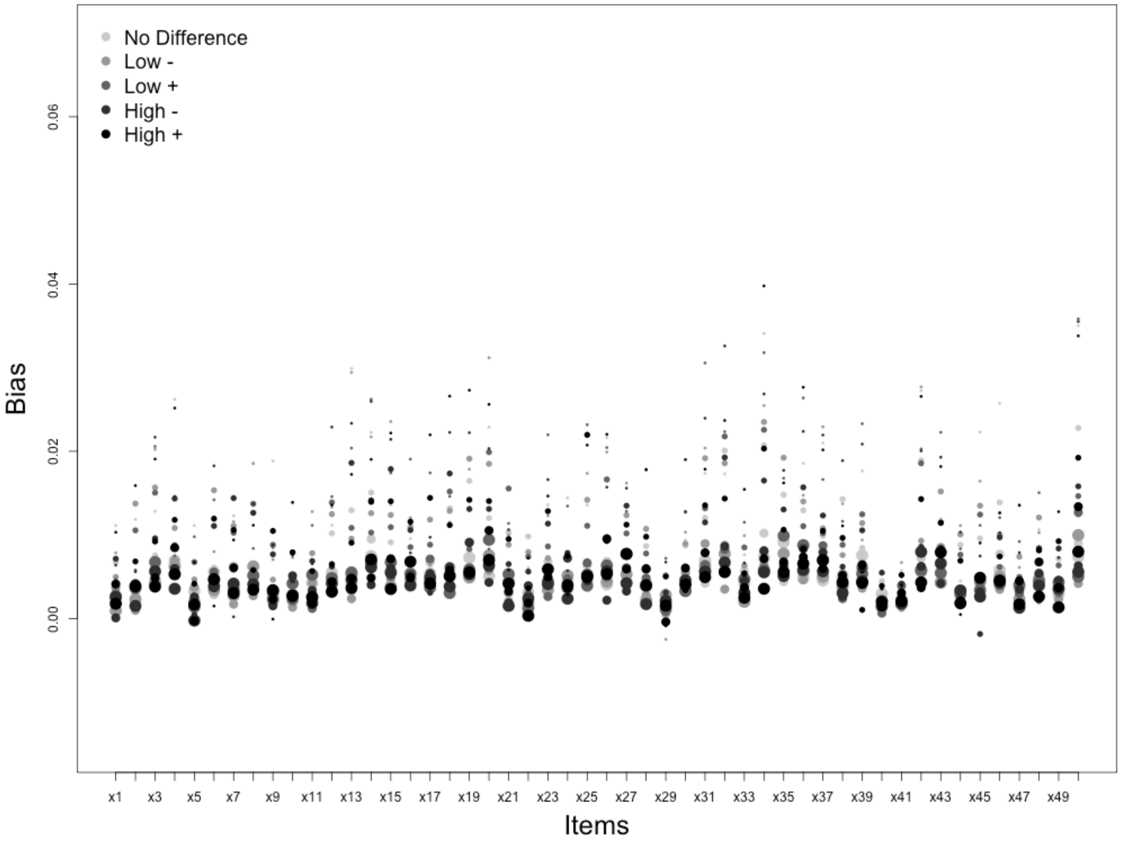

Figure A54: Plot of bias discrimination for the 50-item test under the cut-score scenario.

Note: The sample size is represented by the increase size in dots on the plot with the smallest dot representing N = 500 and the largest representing N = 10,000.

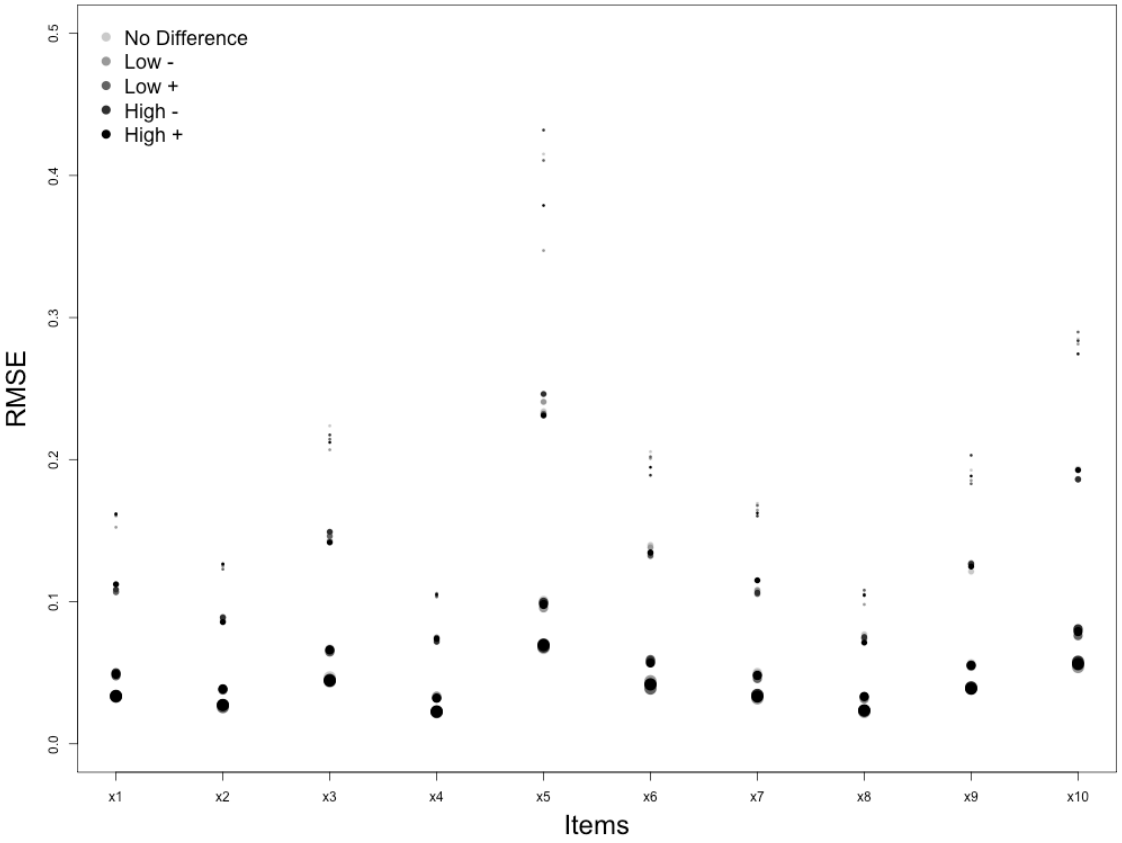

Figure A55: Plot of RMSE difficulty for the 10-item test under the cut-score scenario.

Note: The sample size is represented by the increase size in dots on the plot with the smallest dot representing N = 500 and the largest representing N = 10,000.

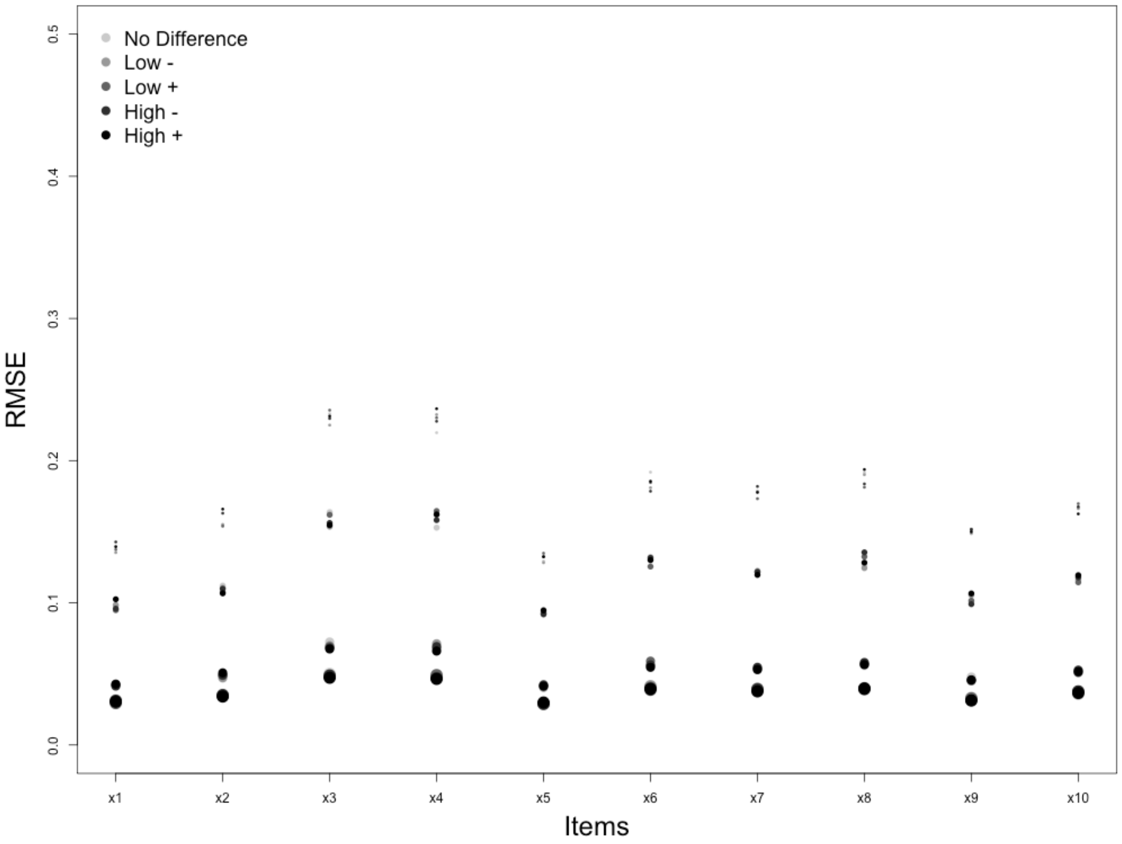

Figure A56: Plot of RMSE discrimination for the 10-item test under the cut-score scenario.

Note: The sample size is represented by the increase size in dots on the plot with the smallest dot representing N = 500 and the largest representing N = 10,000.

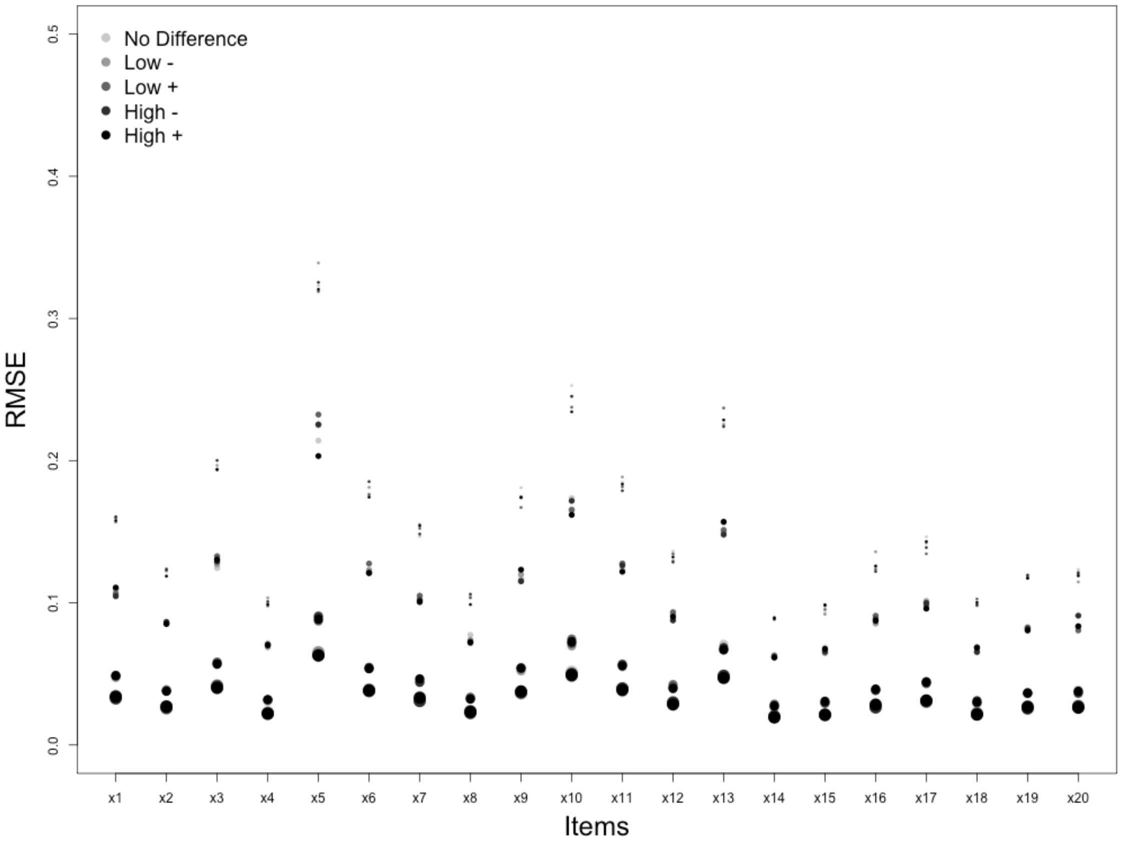

Figure A57: Plot of RMSE difficulty for the 20-item test under the cut-score scenario.

Note: The sample size is represented by the increase size in dots on the plot with the smallest dot representing N = 500 and the largest representing N = 10,000.

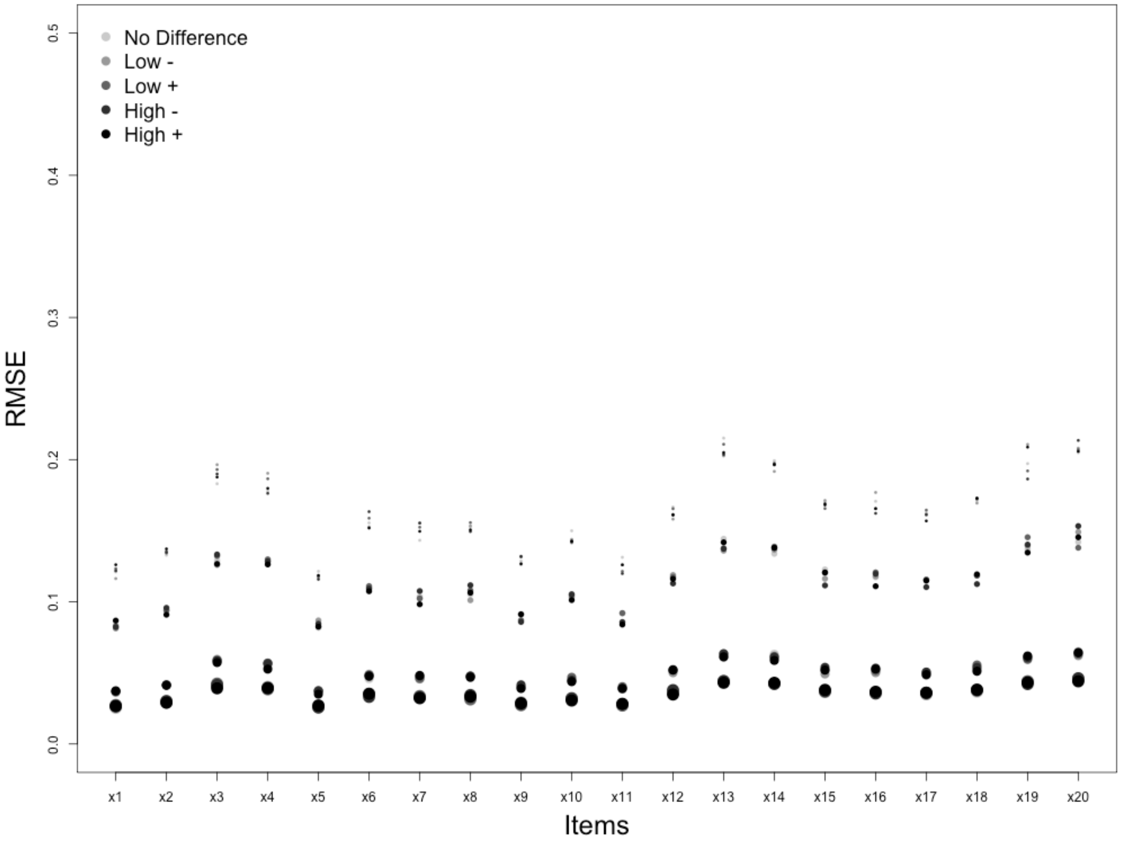

Figure A58: Plot of RMSE discrimination for the 20-item test under the cut-score scenario.

Note: The sample size is represented by the increase size in dots on the plot with the smallest dot representing N = 500 and the largest representing N = 10,000.

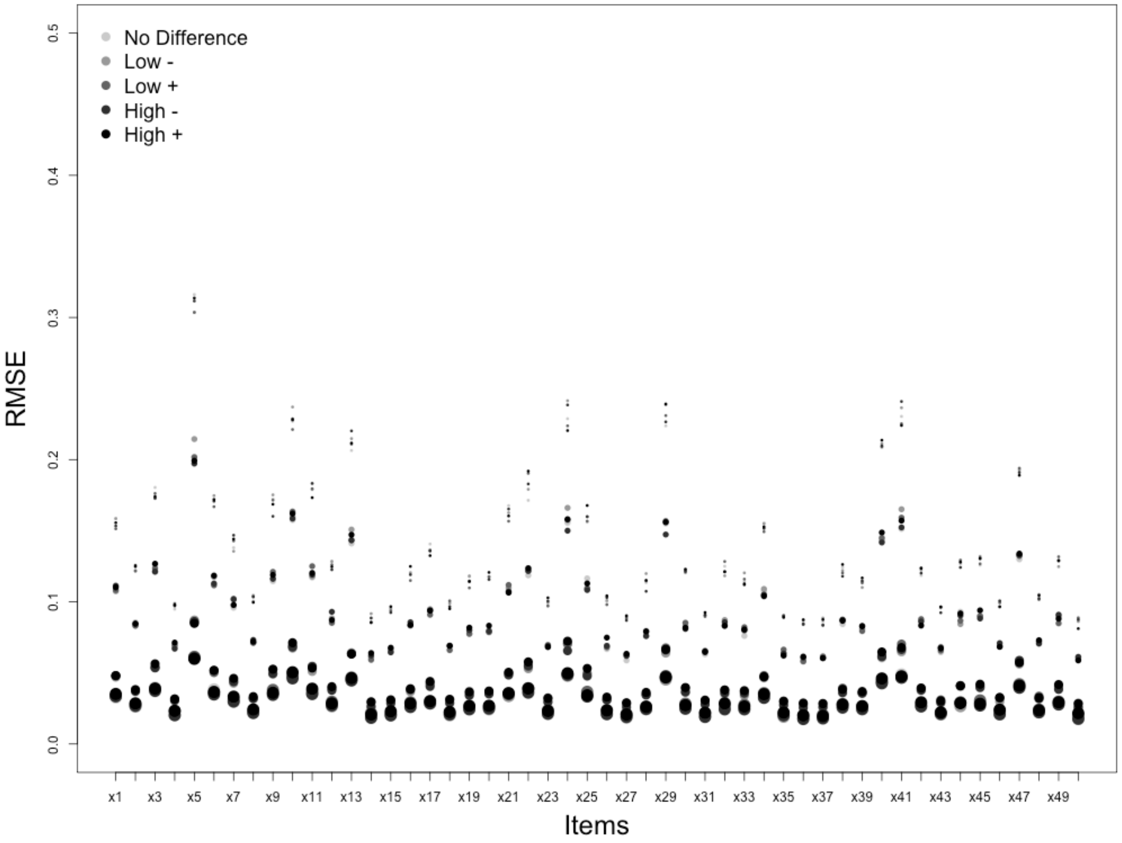

Figure A59: Plot of RMSE difficulty for the 50-item test under the cut-score scenario.

Note: The sample size is represented by the increase size in dots on the plot with the smallest dot representing N = 500 and the largest representing N = 10,000.

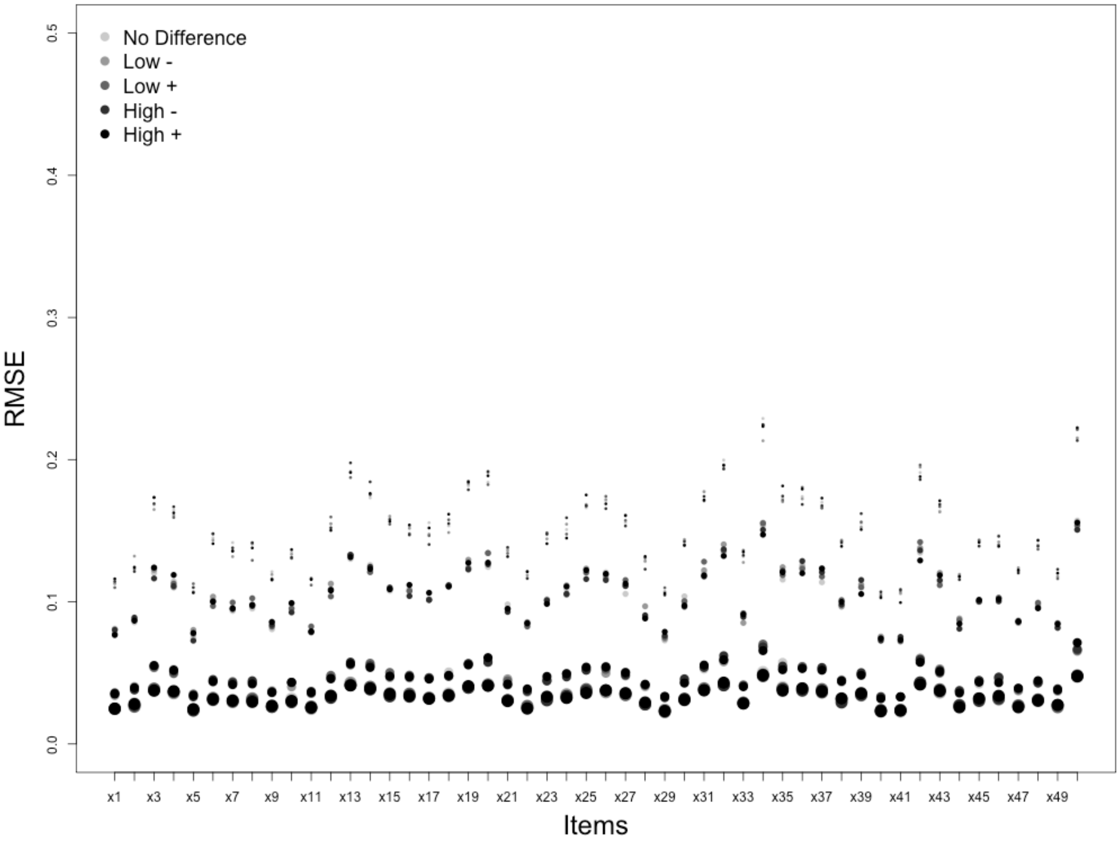

Figure A60: Plot of RMSE discrimination for the 50-item test under the cut-score scenario.

Note: The sample size is represented by the increase size in dots on the plot with the smallest dot representing N = 500 and the largest representing N = 10,000.